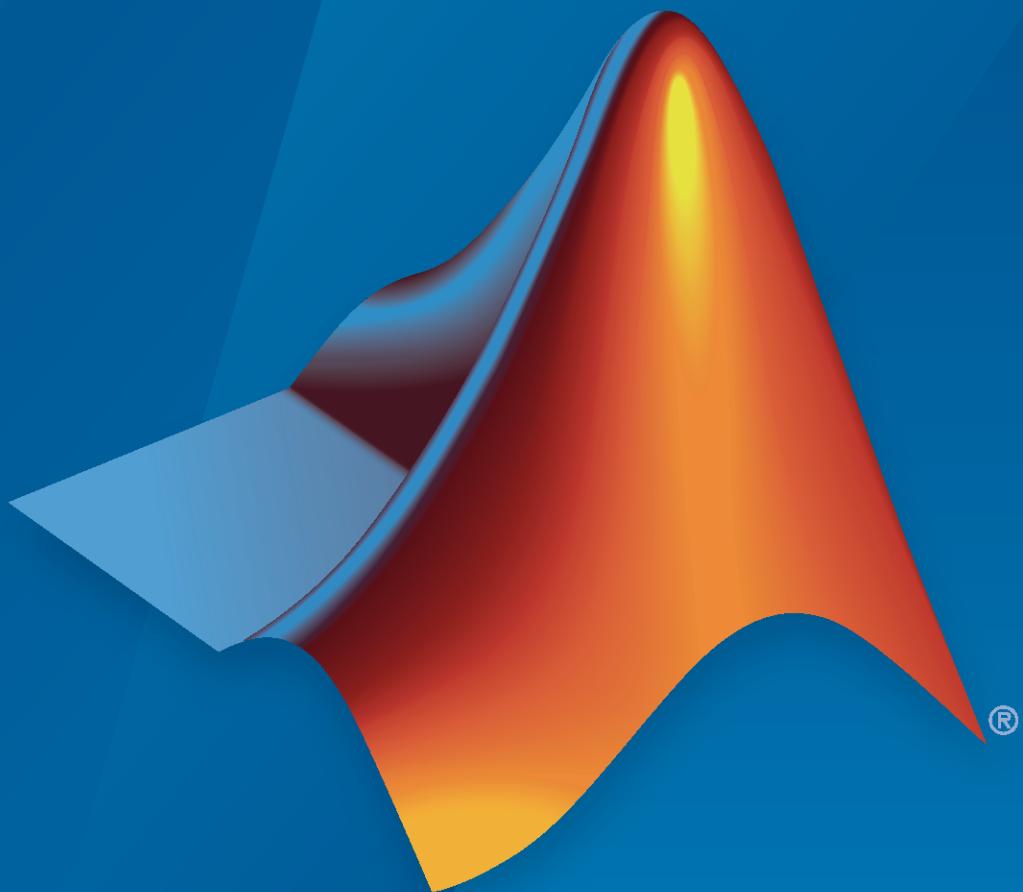


Reinforcement Learning Toolbox™

User's Guide



MATLAB®

R2020a

 MathWorks®

How to Contact MathWorks



Latest news: www.mathworks.com
Sales and services: www.mathworks.com/sales_and_services
User community: www.mathworks.com/matlabcentral
Technical support: www.mathworks.com/support/contact_us
Phone: 508-647-7000



The MathWorks, Inc.
1 Apple Hill Drive
Natick, MA 01760-2098

Reinforcement Learning Toolbox™ User's Guide

© COPYRIGHT 2019- 2020 by The MathWorks, Inc.

The software described in this document is furnished under a license agreement. The software may be used or copied only under the terms of the license agreement. No part of this manual may be photocopied or reproduced in any form without prior written consent from The MathWorks, Inc.

FEDERAL ACQUISITION: This provision applies to all acquisitions of the Program and Documentation by, for, or through the federal government of the United States. By accepting delivery of the Program or Documentation, the government hereby agrees that this software or documentation qualifies as commercial computer software or commercial computer software documentation as such terms are used or defined in FAR 12.212, DFARS Part 227.72, and DFARS 252.227-7014. Accordingly, the terms and conditions of this Agreement and only those rights specified in this Agreement, shall pertain to and govern the use, modification, reproduction, release, performance, display, and disclosure of the Program and Documentation by the federal government (or other entity acquiring for or through the federal government) and shall supersede any conflicting contractual terms or conditions. If this License fails to meet the government's needs or is inconsistent in any respect with federal procurement law, the government agrees to return the Program and Documentation, unused, to The MathWorks, Inc.

Trademarks

MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

Patents

MathWorks products are protected by one or more U.S. patents. Please see www.mathworks.com/patents for more information.

Revision History

March 2019	Online only	New for Version 1.0 (Release 2019a)
September 2019	Online only	Revised for Version 1.1 (Release 2019b)
March 2020	Online only	Revised for Version 1.2 (Release 2020a)

Getting Started

1

Reinforcement Learning Toolbox Product Description	1-2
What Is Reinforcement Learning?	1-3
Reinforcement Learning Workflow	1-4
Reinforcement Learning for Control Systems Applications	1-6
Train Reinforcement Learning Agent in Basic Grid World	1-8
Train Reinforcement Learning Agent in MDP Environment	1-14
Create Simulink Environment and Train Agent	1-19

Create Environments

2

Create MATLAB Environments for Reinforcement Learning	2-2
Action and Observation Signals	2-2
Predefined MATLAB Environments	2-3
Custom MATLAB Environments	2-3
Create Simulink Environments for Reinforcement Learning	2-5
Action and Observation Signals	2-5
Predefined Simulink Environments	2-6
Custom Simulink Environments	2-6
Define Reward Signals	2-7
Continuous Rewards	2-7
Discrete Rewards	2-8
Mixed Rewards	2-8
Load Predefined Grid World Environments	2-9
Basic Grid World	2-9
Deterministic Waterfall Grid Worlds	2-10
Stochastic Waterfall Grid Worlds	2-12
Load Predefined Control System Environments	2-15
Cart-Pole Environments	2-15
Double Integrator Environments	2-17
Simple Pendulum Environments with Image Observation	2-19

Load Predefined Simulink Environments	2-22
Simple Pendulum Simulink Model	2-22
Cart-Pole Simscape Model	2-24
Create Custom Grid World Environments	2-28
Grid World Model	2-28
Grid World Environment	2-32
Create MATLAB Environment Using Custom Functions	2-33
Create Custom MATLAB Environment from Template	2-40
Create Template Class	2-40
Environment Properties	2-40
Required Functions	2-41
Optional Functions	2-43
Environment Visualization	2-44
Create Custom Environment	2-45
Water Tank Reinforcement Learning Environment Model	2-46

Define Policies and Value Functions

3

Create Policy and Value Function Representations	3-2
Function Approximation	3-2
Table Representations	3-3
Deep Neural Network Representations	3-3
Linear Basis Function Representations	3-7
Specify Agent Representations	3-7
Import Policy and Value Function Representations	3-9
Import Actor and Critic for Image Observation Application	3-9

Create Agents

4

Reinforcement Learning Agents	4-2
Built-In Agents	4-3
Custom Agents	4-3
Q-Learning Agents	4-4
Critic Function	4-4
Agent Creation	4-4
Training Algorithm	4-4
SARSA Agents	4-6
Critic Function	4-6
Agent Creation	4-6
Training Algorithm	4-6

Deep Q-Network Agents	4-8
Critic Function	4-8
Agent Creation	4-8
Training Algorithm	4-9
Target Update Methods	4-9
Policy Gradient Agents	4-11
Actor and Critic Functions	4-11
Agent Creation	4-11
Training Algorithm	4-11
Deep Deterministic Policy Gradient Agents	4-14
Actor and Critic Function	4-14
Agent Creation	4-14
Training Algorithm	4-15
Target Update Methods	4-16
Twin-Delayed Deep Deterministic Policy Gradient Agents	4-17
Actor and Critic Function	4-17
Agent Creation	4-18
Training Algorithm	4-18
Target Update Methods	4-19
Actor-Critic Agents	4-21
Actor and Critic Function	4-21
Agent Creation	4-21
Training Algorithm	4-21
Proximal Policy Optimization Agents	4-24
Actor and Critic Function	4-24
Agent Creation	4-24
Training Algorithm	4-25
Custom Agents	4-28
Create Template Class	4-28
Agent Properties	4-28
Constructor Function	4-29
Actor and Critic Representations	4-30
Required Functions	4-30
Optional Functions	4-33
Create Custom Agent	4-33

Train and Validate Agents

5

Train Reinforcement Learning Agents	5-2
Training Algorithm	5-2
Episode Manager	5-3
Save Candidate Agents	5-4
Parallel Computing	5-4
GPU Acceleration	5-6
Validate Trained Policy	5-6

Environment Visualization	5-6
Train DQN Agent to Balance Cart-Pole System	5-8
Train PG Agent to Balance Cart-Pole System	5-14
Train AC Agent to Balance Cart-Pole System	5-19
Train PG Agent with Baseline to Control Double Integrator System ...	5-25
Train DDPG Agent to Control Double Integrator System	5-31
Train DQN Agent to Swing Up and Balance Pendulum	5-37
Train DDPG Agent to Swing Up and Balance Pendulum	5-43
Train DDPG Agent to Swing Up and Balance Cart-Pole System	5-50
Train DDPG Agent to Swing Up and Balance Pendulum with Bus Signal	5-57
Train DDPG Agent to Swing Up and Balance Pendulum with Image Observation	5-64
Create Agent Using Deep Network Designer and Train Using Image Observations	5-72
Train AC Agent to Balance Cart-Pole System Using Parallel Computing	5-84
Train DDPG Agent to Control Flying Robot	5-89
Train PPO Agent to Land Rocket	5-95
Train DDPG Agent for Adaptive Cruise Control	5-102
Train DQN Agent for Lane Keeping Assist	5-110
Train DDPG Agent for Path Following Control	5-118
Train DQN Agent for Lane Keeping Assist Using Parallel Computing ..	5-126
Train Biped Robot to Walk Using Reinforcement Learning Agents ...	5-134
Quadruped Robot Locomotion Using DDPG Agent	5-145
Train Custom LQR Agent	5-153
Imitate MPC Controller for Lane Keep Assist	5-157
Train DDPG Agent with Pretrained Actor Network	5-165
Imitate Nonlinear MPC Controller for Flying Robot	5-174

Deploy Trained Policies**6**

Deploy Trained Reinforcement Learning Policies	6-2
Create Policy Evaluation Function	6-2
Generate Code Using GPU Coder	6-2
Generate Code Using MATLAB Coder	6-3

Getting Started

- “Reinforcement Learning Toolbox Product Description” on page 1-2
- “What Is Reinforcement Learning?” on page 1-3
- “Reinforcement Learning for Control Systems Applications” on page 1-6
- “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8
- “Train Reinforcement Learning Agent in MDP Environment” on page 1-14
- “Create Simulink Environment and Train Agent” on page 1-19

Reinforcement Learning Toolbox Product Description

Design and train policies using reinforcement learning

Reinforcement Learning Toolbox™ provides functions and blocks for training policies using reinforcement learning algorithms including DQN, A2C, and DDPG. You can use these policies to implement controllers and decision-making algorithms for complex systems such as robots and autonomous systems. You can implement the policies using deep neural networks, polynomials, or look-up tables.

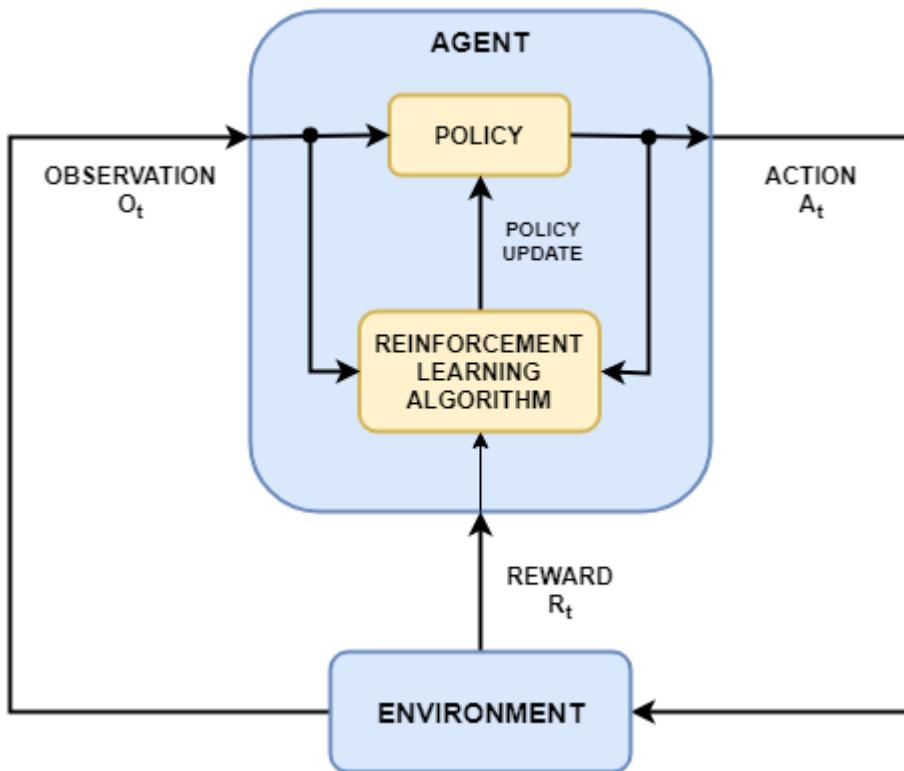
The toolbox lets you train policies by enabling them to interact with environments represented by MATLAB® or Simulink® models. You can evaluate algorithms, experiment with hyperparameter settings, and monitor training progress. To improve training performance, you can run simulations in parallel on the cloud, computer clusters, and GPUs (with Parallel Computing Toolbox™ and MATLAB Parallel Server™).

Through the ONNX™ model format, existing policies can be imported from deep learning frameworks such as TensorFlow™ Keras and PyTorch (with Deep Learning Toolbox™). You can generate optimized C, C++, and CUDA code to deploy trained policies on microcontrollers and GPUs.

The toolbox includes reference examples for using reinforcement learning to design controllers for robotics and automated driving applications.

What Is Reinforcement Learning?

Reinforcement learning is a goal-directed computational approach where a computer learns to perform a task by interacting with an unknown dynamic environment. This learning approach enables a computer to make a series of decisions to maximize the cumulative reward for the task without human intervention and without being explicitly programmed to achieve the task. The following diagram shows a general representation of a reinforcement learning scenario.



The goal of reinforcement learning is to train an *agent* to complete a task within an unknown *environment*. The agent receives *observations* and a *reward* from the environment and sends *actions* to the environment. The reward is a measure of how successful an action is with respect to completing the task goal.

The agent contains two components: a *policy* and a *learning algorithm*.

- The policy is a mapping that selects actions based on the observations from the environment. Typically, the policy is a function approximator with tunable parameters, such as a deep neural network.
- The learning algorithm continuously updates the policy parameters based on the actions, observations, and reward. The goal of the learning algorithm is to find an optimal policy that maximizes the cumulative reward received during the task.

In other words, reinforcement learning involves an agent learning the optimal behavior through repeated trial-and-error interactions with the environment without human involvement.

As an example, consider the task of parking a vehicle using an automated driving system. The goal of this task is for the vehicle computer (**agent**) to park the vehicle in the correct position and

orientation. To do so, the controller uses readings from cameras, accelerometers, gyroscopes, a GPS receiver, and lidar (**observations**) to generate steering, braking, and acceleration commands (**actions**). The action commands are sent to the actuators that control the vehicle. The resulting observations depend on the actuators, sensors, vehicle dynamics, road surface, wind, and many other less-important factors. All these factors, that is, everything that is not the agent, make up the **environment** in reinforcement learning.

To learn how to generate the correct actions from the observations, the computer repeatedly tries to park the vehicle using a trial-and-error process. To guide the learning process, you provide a signal that is one when the car successfully reaches the desired position and orientation and zero otherwise (**reward**). During each trial, the computer selects actions using a mapping (**policy**) initialized with some default values. After each trial, the computer updates the mapping to maximize the reward (**learning algorithm**). This process continues until the computer learns an optimal mapping that successfully parks the car.

Reinforcement Learning Workflow

The general workflow for training an agent using reinforcement learning includes the following steps.



- 1 Formulate Problem** — Define the task for the agent to learn, including how the agent interacts with the environment and any primary and secondary goals the agent must achieve.
- 2 Create Environment** — Define the environment within which the agent operates, including the interface between agent and environment and the environment dynamic model. For more information, see “Create MATLAB Environments for Reinforcement Learning” on page 2-2 and “Create Simulink Environments for Reinforcement Learning” on page 2-5.
- 3 Define Reward** — Specify the reward signal that the agent uses to measure its performance against the task goals and how this signal is calculated from the environment. For more information, see “Define Reward Signals” on page 2-7.
- 4 Create Agent** — Create the agent, which includes defining a policy representation and configuring the agent learning algorithm. For more information, see “Create Policy and Value Function Representations” on page 3-2 and “Reinforcement Learning Agents” on page 4-2.
- 5 Train Agent** — Train the agent policy representation using the defined environment, reward, and agent learning algorithm. For more information, see “Train Reinforcement Learning Agents” on page 5-2.
- 6 Validate Agent** — Evaluate the performance of the trained agent by simulating the agent and environment together. For more information, see “Train Reinforcement Learning Agents” on page 5-2.

-
- 7 Deploy Policy** — Deploy the trained policy representation using, for example, generated GPU code. For more information, see “Deploy Trained Reinforcement Learning Policies” on page 6-2.

Training an agent using reinforcement learning is an iterative process. Decisions and results in later stages can require you to return to an earlier stage in the learning workflow. For example, if the training process does not converge to an optimal policy within a reasonable amount of time, you might have to update any of the following before retraining the agent:

- Training settings
- Learning algorithm configuration
- Policy representation
- Reward signal definition
- Action and observation signals
- Environment dynamics

See Also

More About

- “Reinforcement Learning for Control Systems Applications” on page 1-6
- “Create Simulink Environment and Train Agent” on page 1-19

Reinforcement Learning for Control Systems Applications

The behavior of a reinforcement learning policy—that is, how the policy observes the environment and generates actions to complete a task in an optimal manner—is similar to the operation of a controller in a control system. Reinforcement learning can be translated to a control system representation using the following mapping.



Reinforcement Learning	Control Systems
Policy	Controller

Reinforcement Learning	Control Systems
Environment	<p>Everything that is not the controller — In the preceding diagram, the environment includes the plant, the reference signal, and the calculation of the error. In general, the environment can also include additional elements, such as:</p> <ul style="list-style-type: none"> • Measurement noise • Disturbance signals • Filters • Analog-to-digital and digital-to-analog converters
Observation	<p>Any measurable value from the environment that is visible to the agent — In the preceding diagram, the controller can see the error signal from the environment. You can also create agents that observe, for example, the reference signal, measurement signal, and measurement signal rate of change.</p>
Action	Manipulated variables or control actions
Reward	Function of the measurement, error signal, or some other performance metric — For example, you can implement reward functions that minimize the steady-state error while minimizing control effort.
Learning Algorithm	Adaptation mechanism of an adaptive controller

Many control problems encountered in areas such as robotics and automated driving require complex, nonlinear control architectures. Techniques such as gain scheduling, robust control, and nonlinear model predictive control (MPC) can be used for these problems, but often require significant domain expertise from the control engineer. For example, gains and parameters are difficult to tune. The resulting controllers can pose implementation challenges, such as the computational intensity of nonlinear MPC.

You can use deep neural networks, trained using reinforcement learning, to implement such complex controllers. These systems can be self-taught without intervention from an expert control engineer. Also, once the system is trained, you can deploy the reinforcement learning policy in a computationally efficient way.

You can also use reinforcement learning to create an end-to-end controller that generates actions directly from raw data, such as images. This approach is attractive for video-intensive applications, such as automated driving, since you do not have to manually define and select image features.

See Also

More About

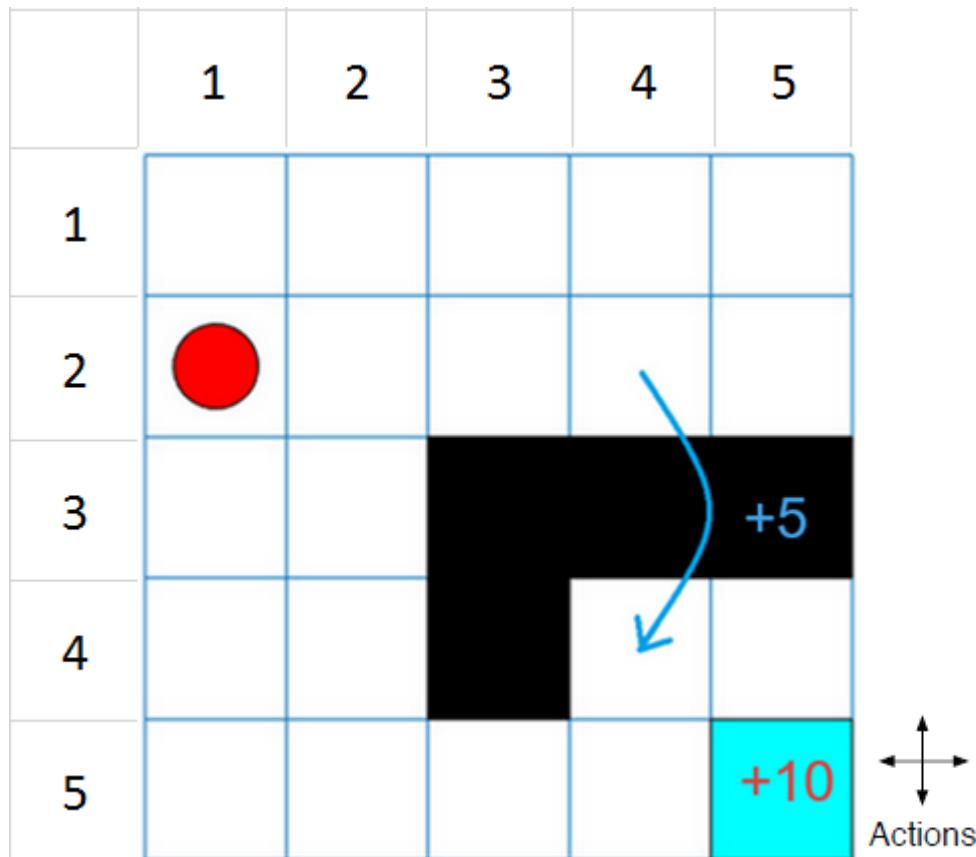
- “What Is Reinforcement Learning?” on page 1-3
- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Define Reward Signals” on page 2-7

Train Reinforcement Learning Agent in Basic Grid World

This example shows how to solve a grid world environment using reinforcement learning by training Q-learning and SARSA agents. For more information on these agents, see “Q-Learning Agents” on page 4-4 and “SARSA Agents” on page 4-6.

This grid world environment has the following configuration and rules:

- 1 The grid world is 5-by-5 and bounded by borders, with four possible actions (North = 1, South = 2, East = 3, West = 4).
- 2 The agent begins from cell [2,1] (second row, first column).
- 3 The agent receives a reward +10 if it reaches the terminal state at cell [5,5] (blue).
- 4 The environment contains a special jump from cell [2,4] to cell [4,4] with a reward of +5.
- 5 The agent is blocked by obstacles (black cells).
- 6 All other actions result in -1 reward.



Create Grid World Environment

Create the basic grid world environment.

```
env = rlPredefinedEnv("BasicGridWorld");
```

To specify that the initial state of the agent is always [2,1], create a reset function that returns the state number for the initial agent state. This function is called at the start of each training episode

and simulation. States are numbered starting at position [1,1]. The state number increases as you move down the first column and then down each subsequent column. Therefore, create an anonymous function handle that sets the initial state to 2.

```
env.ResetFcn = @( ) 2;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create Q-Learning Agent

To create a Q-learning agent, first create a Q table using the observation and action specifications from the grid world environment. Set the learning rate of the representation to 1.

```
qTable = rlTable(getObservationInfo(env),getActionInfo(env));
qRepresentation = rlQValueRepresentation(qTable,getObservationInfo(env),getActionInfo(env));
qRepresentation.Options.LearnRate = 1;
```

Next, create a Q-learning agent using this table representation and configure the epsilon-greedy exploration. For more information on creating Q-learning agents, see `rlQAgent` and `rlQAgentOptions`.

```
agentOpts = rlQAgentOptions;
agentOpts.EpsilonGreedyExploration.Epsilon = .04;
qAgent = rlQAgent(qRepresentation,agentOpts);
```

Train Q-Learning Agent

To train the agent, first specify the training options. For this example, use the following options:

- Train for at most 200 episodes. Specify that each episode lasts for most 50 time steps.
- Stop training when the agent receives an average cumulative reward greater than 10 over 30 consecutive episodes.

For more information, see `rlTrainingOptions`.

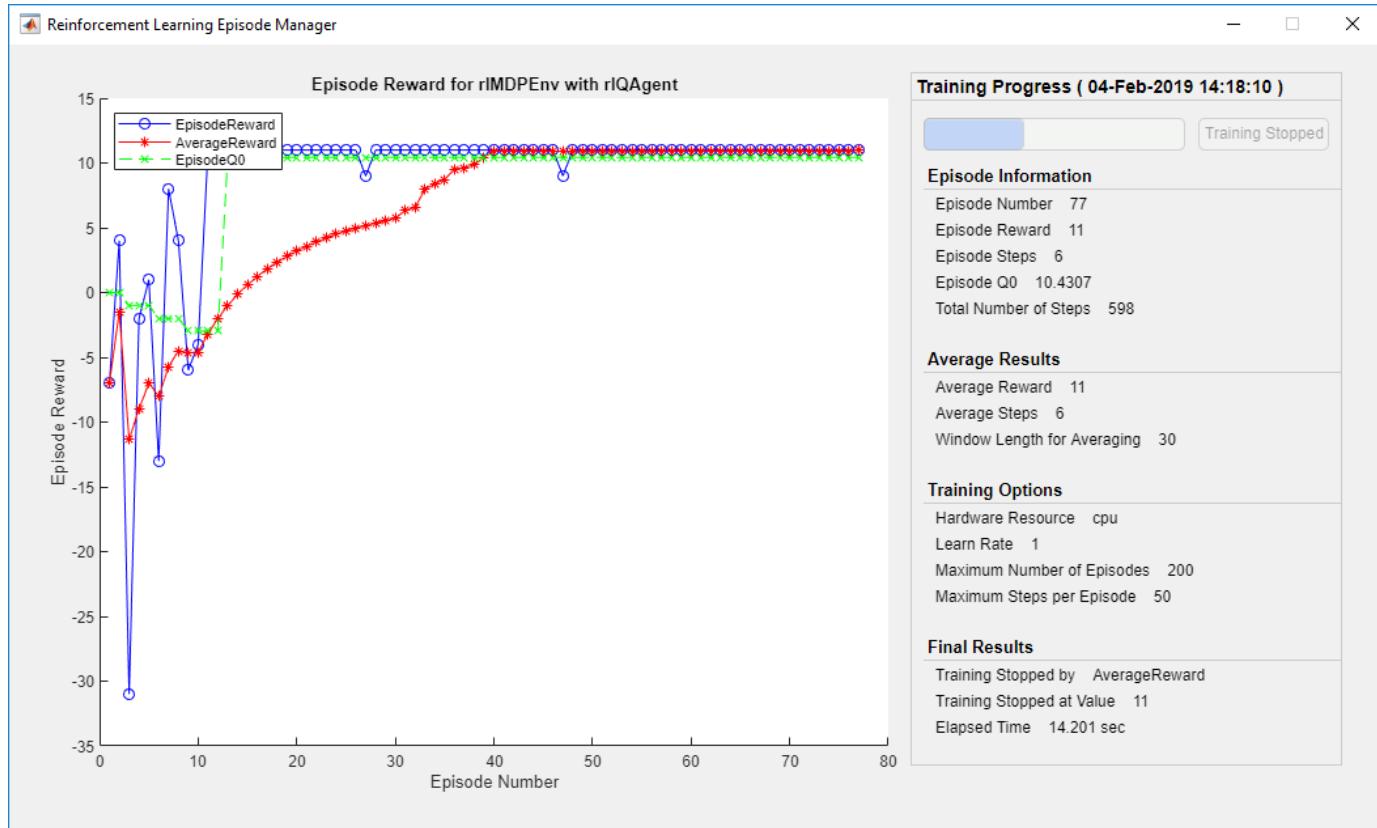
```
trainOpts = rlTrainingOptions;
trainOpts.MaxStepsPerEpisode = 50;
trainOpts.MaxEpisodes= 200;
trainOpts.StopTrainingCriteria = "AverageReward";
trainOpts.StopTrainingValue = 11;
trainOpts.ScoreAveragingWindowLength = 30;
```

Train the Q-learning agent using the `train` function. Training can take several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(qAgent,env,trainOpts);
else
    % Load the pretrained agent for the example.
    load('basicGWQAgent.mat','qAgent')
end
```

The **Episode Manager** window opens and displays the training progress.



Validate Q-Learning Results

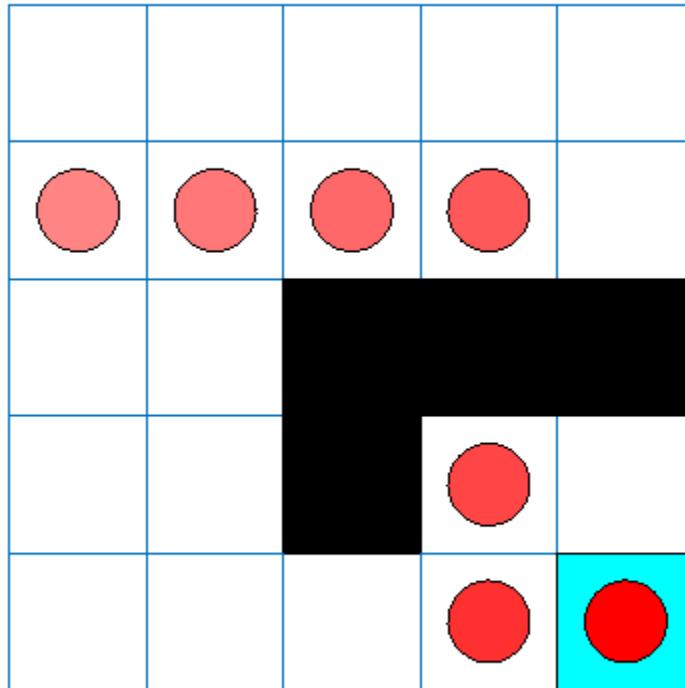
To validate the training results, simulate the agent in the training environment.

Before running the simulation, visualize the environment and configure the visualization to maintain a trace of the agent states.

```
plot(env)
env.Model.Viewer.ShowTrace = true;
env.Model.Viewer.clearTrace;
```

Simulate the agent in the environment using the `sim` function.

```
sim(qAgent,env)
```



The agent trace shows that the agent successfully finds the jump from cell [2,4] to cell [4,4].

Create and Train SARSA Agent

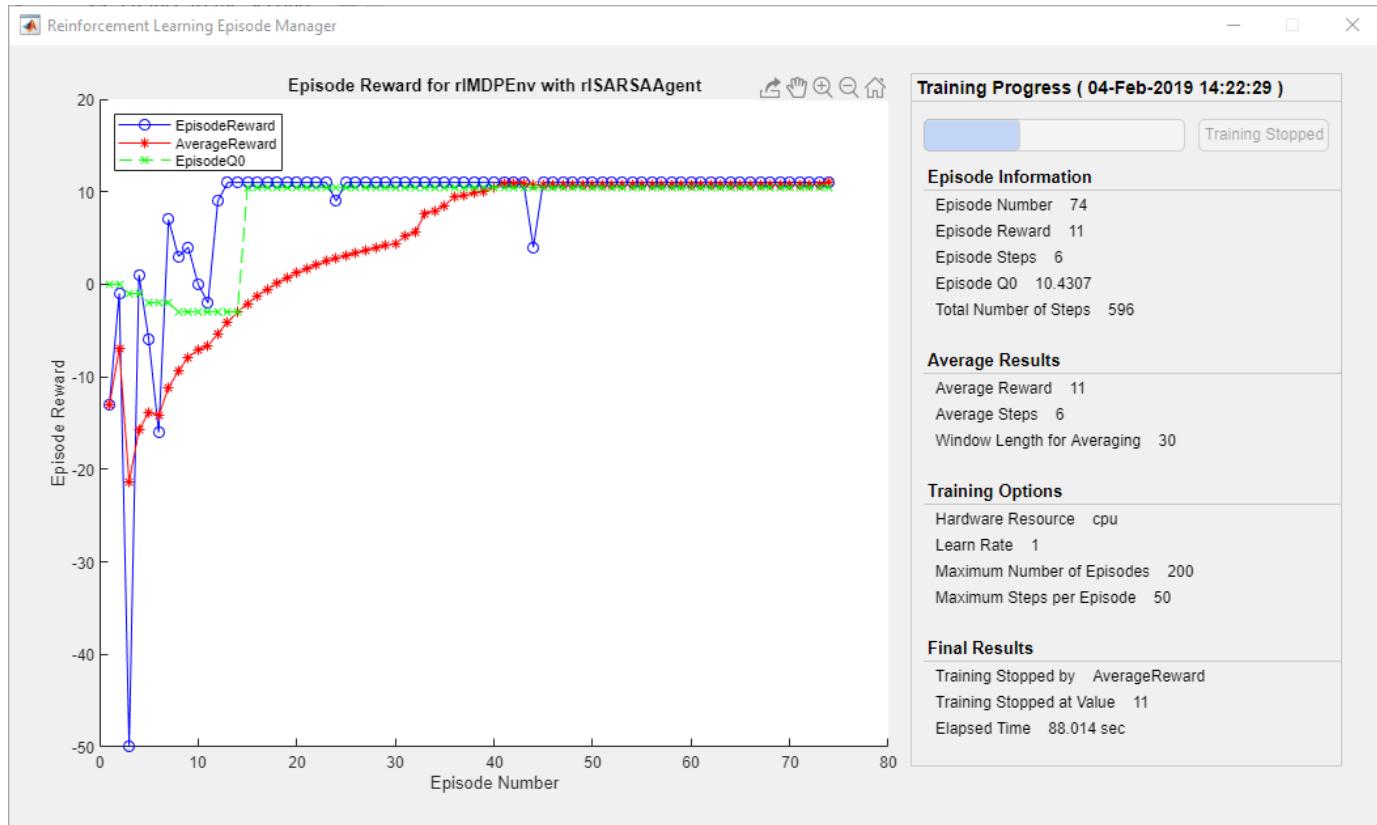
To create a SARSA agent, use the same Q table representation and epsilon-greedy configuration as for the Q-learning agent. For more information on creating SARSA agents, see `rlSARSAAgent` and `rlSARSAOptions`.

```
agentOpts = rlSARSAOptions;
agentOpts.EpsilonGreedyExploration.Epsilon = 0.04;
sarsaAgent = rlSARSAAgent(qRepresentation,agentOpts);
```

Train the SARSA agent using the `train` function. Training can take several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(sarsaAgent,env,trainOpts);
else
    % Load the pretrained agent for the example.
    load('basicGWSarsaAgent.mat','sarsaAgent')
end
```



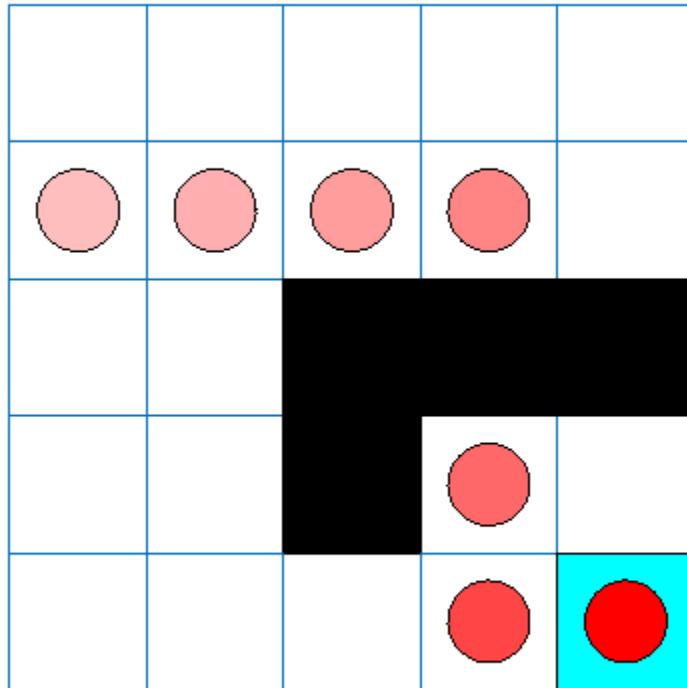
Validate SARSA Training

To validate the training results, simulate the agent in the training environment.

```
plot(env)
env.Model.Viewer.ShowTrace = true;
env.Model.Viewer.clearTrace;
```

Simulate the agent in the environment.

```
sim(sarsaAgent, env)
```



The SARSA agent finds the same grid world solution as the Q-learning agent.

See Also

`createGridWorld` | `rlMDPEnv`

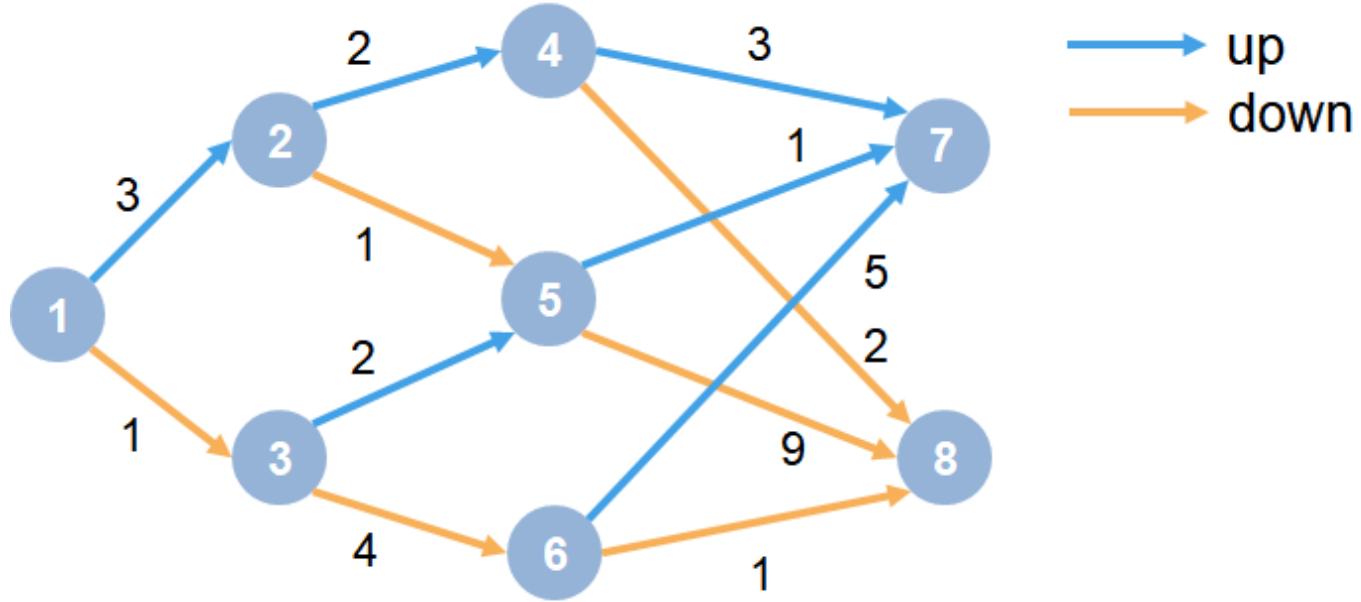
More About

- “Reinforcement Learning Agents” on page 4-2
- “Train Reinforcement Learning Agents” on page 5-2

Train Reinforcement Learning Agent in MDP Environment

This example shows how to train a Q-learning agent to solve a generic Markov decision process (MDP) environment. For more information on these agents, see "Q-Learning Agents" on page 4-4.

The MDP environment has the following graph.



Here:

- 1 Each circle represents a state.
- 2 At each state there is a decision to go up or down.
- 3 The agent begins from state 1.
- 4 The agent receives a reward equal to the value on each transition in the graph.
- 5 The training goal is to collect the maximum cumulative reward.

Create MDP Environment

Create an MDP model with eight states and two actions ("up" and "down").

```
MDP = createMDP(8, ["up"; "down"]);
```

To model the transitions from the above graph, modify the state transition matrix and reward matrix of the MDP. By default, these matrices contain zeros. For more information on creating an MDP model and the properties of an MDP object, see `createMDP`.

Specify the state transition and reward matrices for the MDP. For example, in the following commands:

- The first two lines specify the transition from state 1 to state 2 by taking action 1 ("up") and a reward of +3 for this transition.
- The next two lines specify the transition from state 1 to state 3 by taking action 2 ("down") and a reward of +1 for this transition.

```
MDP.T(1,2,1) = 1;
MDP.R(1,2,1) = 3;
MDP.T(1,3,2) = 1;
MDP.R(1,3,2) = 1;
```

Similarly, specify the state transitions and rewards for the remaining rules in the graph.

```
% State 2 transition and reward
MDP.T(2,4,1) = 1;
MDP.R(2,4,1) = 2;
MDP.T(2,5,2) = 1;
MDP.R(2,5,2) = 1;
% State 3 transition and reward
MDP.T(3,5,1) = 1;
MDP.R(3,5,1) = 2;
MDP.T(3,6,2) = 1;
MDP.R(3,6,2) = 4;
% State 4 transition and reward
MDP.T(4,7,1) = 1;
MDP.R(4,7,1) = 3;
MDP.T(4,8,2) = 1;
MDP.R(4,8,2) = 2;
% State 5 transition and reward
MDP.T(5,7,1) = 1;
MDP.R(5,7,1) = 1;
MDP.T(5,8,2) = 1;
MDP.R(5,8,2) = 9;
% State 6 transition and reward
MDP.T(6,7,1) = 1;
MDP.R(6,7,1) = 5;
MDP.T(6,8,2) = 1;
MDP.R(6,8,2) = 1;
% State 7 transition and reward
MDP.T(7,7,1) = 1;
MDP.R(7,7,1) = 0;
MDP.T(7,7,2) = 1;
MDP.R(7,7,2) = 0;
% State 8 transition and reward
MDP.T(8,8,1) = 1;
MDP.R(8,8,1) = 0;
MDP.T(8,8,2) = 1;
MDP.R(8,8,2) = 0;
```

Specify states "s7" and "s8" as terminal states of the MDP.

```
MDP.TerminalStates = ["s7"; "s8"];
```

Create the reinforcement learning MDP environment for this process model.

```
env = rlMDPEnv(MDP);
```

To specify that the initial state of the agent is always state 1, specify a reset function that returns the initial agent state. This function is called at the start of each training episode and simulation. Create an anonymous function handle that sets the initial state to 1.

```
env.ResetFcn = @() 1;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create Q-Learning Agent

To create a Q-learning agent, first create a Q table using the observation and action specifications from the MDP environment. Set the learning rate of the representation to 1.

```
obsInfo = getObservationInfo(env);  
actInfo = getActionInfo(env);  
qTable = rlTable(obsInfo, actInfo);  
qRepresentation = rlValueRepresentation(qTable, obsInfo, actInfo);  
qRepresentation.Options.LearnRate = 1;
```

Next, create a Q-learning agent using this table representation, configuring the epsilon-greedy exploration. For more information on creating Q-learning agents, see `rlQAgent` and `rlQAgentOptions`.

```
agentOpts = rlQAgentOptions;  
agentOpts.DiscountFactor = 1;  
agentOpts.EpsilonGreedyExploration.Epsilon = 0.9;  
agentOpts.EpsilonGreedyExploration.EpsilonDecay = 0.01;  
qAgent = rlQAgent(qRepresentation, agentOpts);
```

Train Q-Learning Agent

To train the agent, first specify the training options. For this example, use the following options:

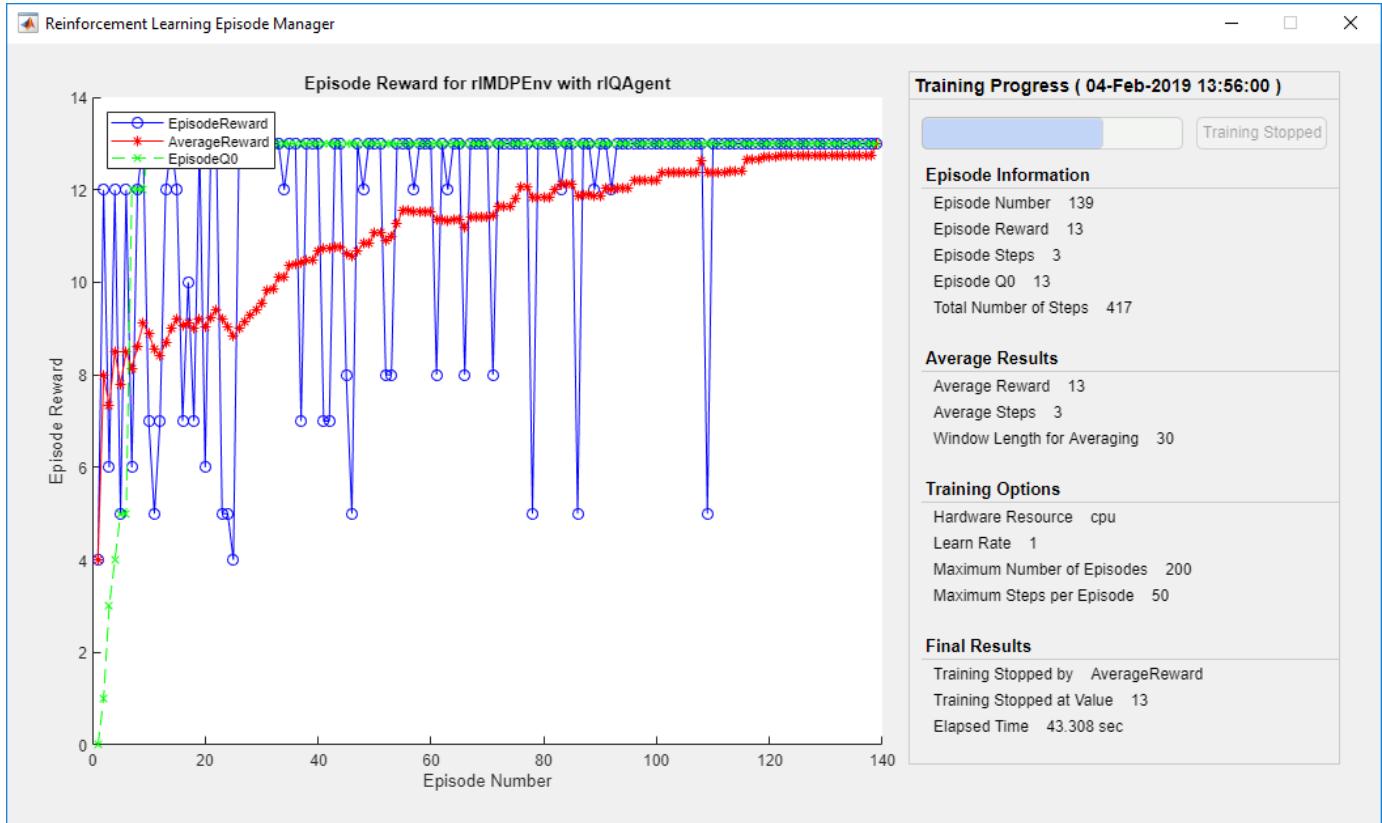
- Train for at most 200 episodes, with each episode lasting at most 50 time steps.
- Stop training when the agent receives an average cumulative reward greater than 10 over 30 consecutive episodes.

For more information, see `rlTrainingOptions`.

```
trainOpts = rlTrainingOptions;  
trainOpts.MaxStepsPerEpisode = 50;  
trainOpts.MaxEpisodes = 200;  
trainOpts.StopTrainingCriteria = "AverageReward";  
trainOpts.StopTrainingValue = 13;  
trainOpts.ScoreAveragingWindowLength = 30;
```

Train the agent using the `train` function. This may take several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;  
  
if doTraining  
    % Train the agent.  
    trainingStats = train(qAgent, env, trainOpts);  
else  
    % Load pretrained agent for the example.  
    load('genericMDPQAgent.mat', 'qAgent');  
end
```



Validate Q-Learning Results

To validate the training results, simulate the agent in the training environment using the `sim` function. The agent successfully finds the optimal path which results in cumulative reward of 13.

```
Data = sim(qAgent,env);
cumulativeReward = sum(Data.Reward)

cumulativeReward = 13
```

Since the discount factor is set to 1, the values in the Q table of the trained agent match the undiscounted returns of the environment.

```
QTable = getLearnableParameters(getCritic(qAgent));
QTable{1}

ans = 8x2

13    12
 5    10
11     9
 3     2
 1     9
 5     1
 0     0
 0     0
```

```
TrueTableValues = [13,12;5,10;11,9;3,2;1,9;5,1;0,0;0,0]
```

```
TrueTableValues = 8x2
```

```
13    12
 5    10
11    9
 3    2
 1    9
 5    1
 0    0
 0    0
```

See Also

[createMDP](#) | [rlMDPEnv](#)

More About

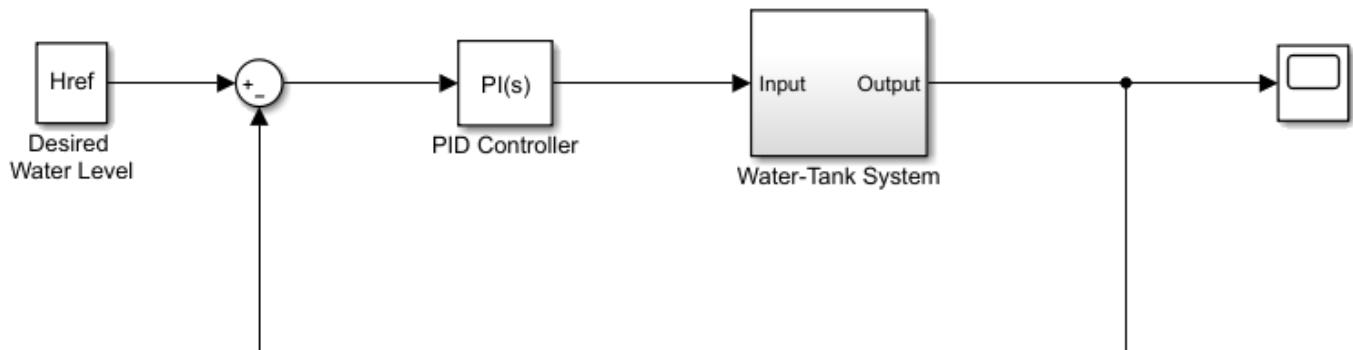
- “Reinforcement Learning Agents” on page 4-2
- “Train Reinforcement Learning Agents” on page 5-2

Create Simulink Environment and Train Agent

This example shows how to convert the PI controller in the `watertank` Simulink® model, to a reinforcement learning deep deterministic policy gradient (DDPG) agent. For an example that trains a DDPG agent in MATLAB®, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Water Tank Model

The original model for this example is the water tank model. The goal is to control the level of the water in the tank. For more information about the water tank model, see “`watertank` Simulink Model” (Simulink Control Design).

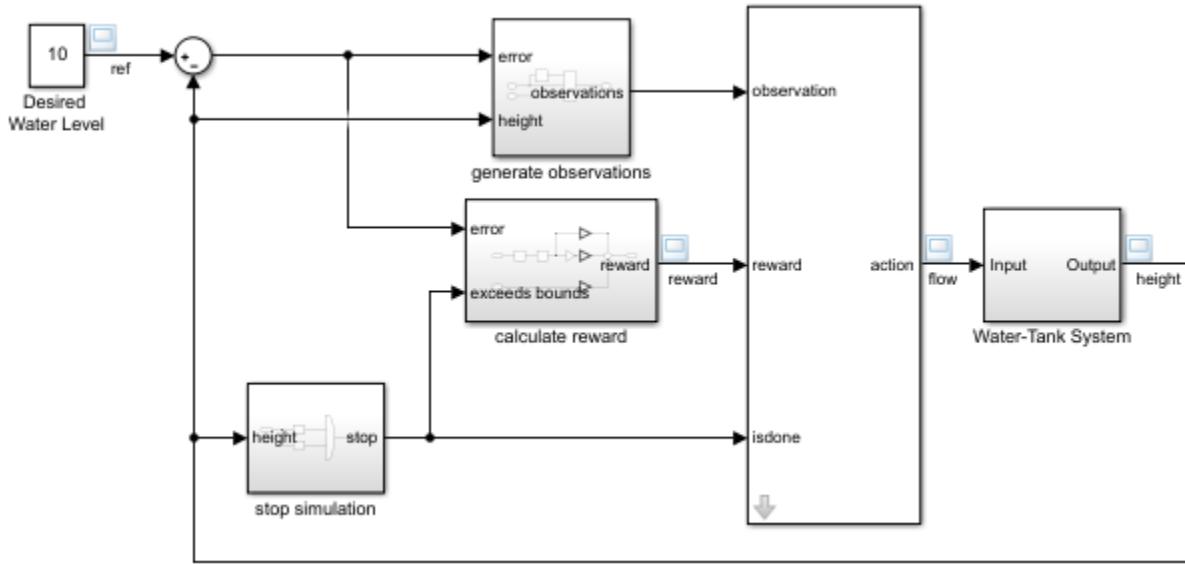


Modify the original model by making the following changes:

- 1** Delete the PID Controller.
- 2** Insert the RL Agent block.
- 3** Connect the observation vector $[e \ dt \ e \ h]^T$, where h is the height of the tank, $e = r - h$, and r is the reference height.
- 4** Set up the reward $\text{reward} = 10(|e| < 0.1) - 1(|e| \geq 0.1) - 100(h \leq 0 \mid h \geq 20)$.
- 5** Configure the termination signal such that the simulation stops if $h \leq 0$ or $h \geq 20$.

The resulting model is `rlwatertank.slx`. For more information on this model and the changes, see “Create Simulink Environments for Reinforcement Learning” on page 2-5.

```
open_system('rlwatertank')
```



Create Environment Interface

Creating an environment model includes defining the following:

- Action and observation signals that the agent uses to interact with the environment. For more information, see [rlNumericSpec](#) and [rlFiniteSetSpec](#).
- Reward signal that the agent uses to measure its success. For more information, see “[Define Reward Signals](#)” on page 2-7.

Define the observation specification `obsInfo` and action specification `actInfo`.

```
obsInfo = rlNumericSpec([3 1],...
    'LowerLimit',[-inf -inf 0 ],...
    'UpperLimit',[ inf inf inf]);
obsInfo.Name = 'observations';
obsInfo.Description = 'integrated error, error, and measured height';
numObservations = obsInfo.Dimension(1);

actInfo = rlNumericSpec([1 1]);
actInfo.Name = 'flow';
numActions = actInfo.Dimension(1);
```

Build the environment interface object.

```
env = rlSimulinkEnv('rlwatertank','rlwatertank/RL Agent',...
    obsInfo,actInfo);
```

Set a custom reset function that randomizes the reference values for the model.

```
env.ResetFcn = @(in)localResetFcn(in);
```

Specify the simulation time `Tf` and the agent sample time `Ts` in seconds.

```
Ts = 1.0;
Tf = 200;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

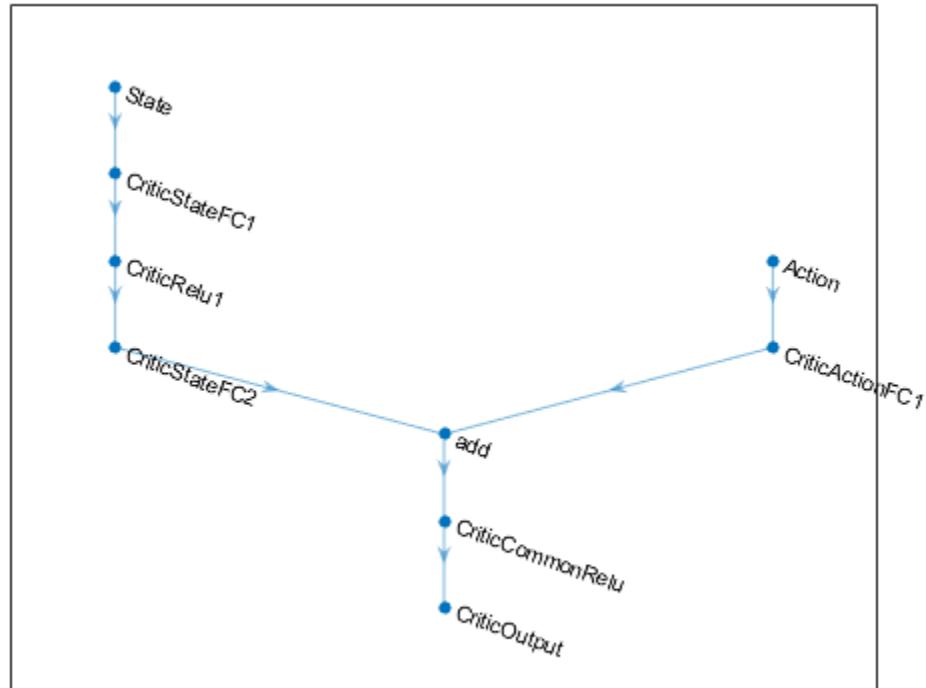
Create DDPG Agent

Given observations and actions, a DDPG agent approximates the long-term reward using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the observation and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
statePath = [  
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'State')  
    fullyConnectedLayer(50, 'Name', 'CriticStateFC1')  
    reluLayer('Name', 'CriticRelu1')  
    fullyConnectedLayer(25, 'Name', 'CriticStateFC2')];  
actionPath = [  
    imageInputLayer([numActions 1 1], 'Normalization', 'none', 'Name', 'Action')  
    fullyConnectedLayer(25, 'Name', 'CriticActionFC1')];  
commonPath = [  
    additionLayer(2, 'Name', 'add')  
    reluLayer('Name', 'CriticCommonRelu')  
    fullyConnectedLayer(1, 'Name', 'CriticOutput')];  
  
criticNetwork = layerGraph();  
criticNetwork = addLayers(criticNetwork, statePath);  
criticNetwork = addLayers(criticNetwork, actionPath);  
criticNetwork = addLayers(criticNetwork, commonPath);  
criticNetwork = connectLayers(criticNetwork, 'CriticStateFC2', 'add/in1');  
criticNetwork = connectLayers(criticNetwork, 'CriticActionFC1', 'add/in2');
```

View the critic network configuration.

```
figure  
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation specifications for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,'Observation',{'State'},{'Action'},{});
```

Given observations, a DDPG agent decides which action to take using an actor representation. To create the actor, first create a deep neural network with one input, the observation, and one output, the action.

Construct the actor in a similar manner to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization','none','Name','State')
    fullyConnectedLayer(3, 'Name','actorFC')
    tanhLayer('Name','actorTanh')
    fullyConnectedLayer(numActions,'Name','Action')
];
actorOptions = rlRepresentationOptions('LearnRate',1e-04,'GradientThreshold',1);
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,'Observation',{'State'},{});
```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```
agentOpts = rlDDPGAgentOptions( ...
    'SampleTime', Ts, ...
    'TargetSmoothFactor', 1e-3, ...
    'DiscountFactor', 1.0, ...
    'MiniBatchSize', 64, ...
    'ExperienceBufferLength', 1e6);
agentOpts.NoiseOptions.Variance = 0.3;
agentOpts.NoiseOptions.VarianceDecayRate = 1e-5;
```

Then, create the DDPG agent using the specified actor representation, critic representation, and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor, critic, agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training for at most 5000 episodes. Specify that each episode lasts for at most 200 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than 800 over 20 consecutive episodes. At this point, the agent can control the level of water in the tank.

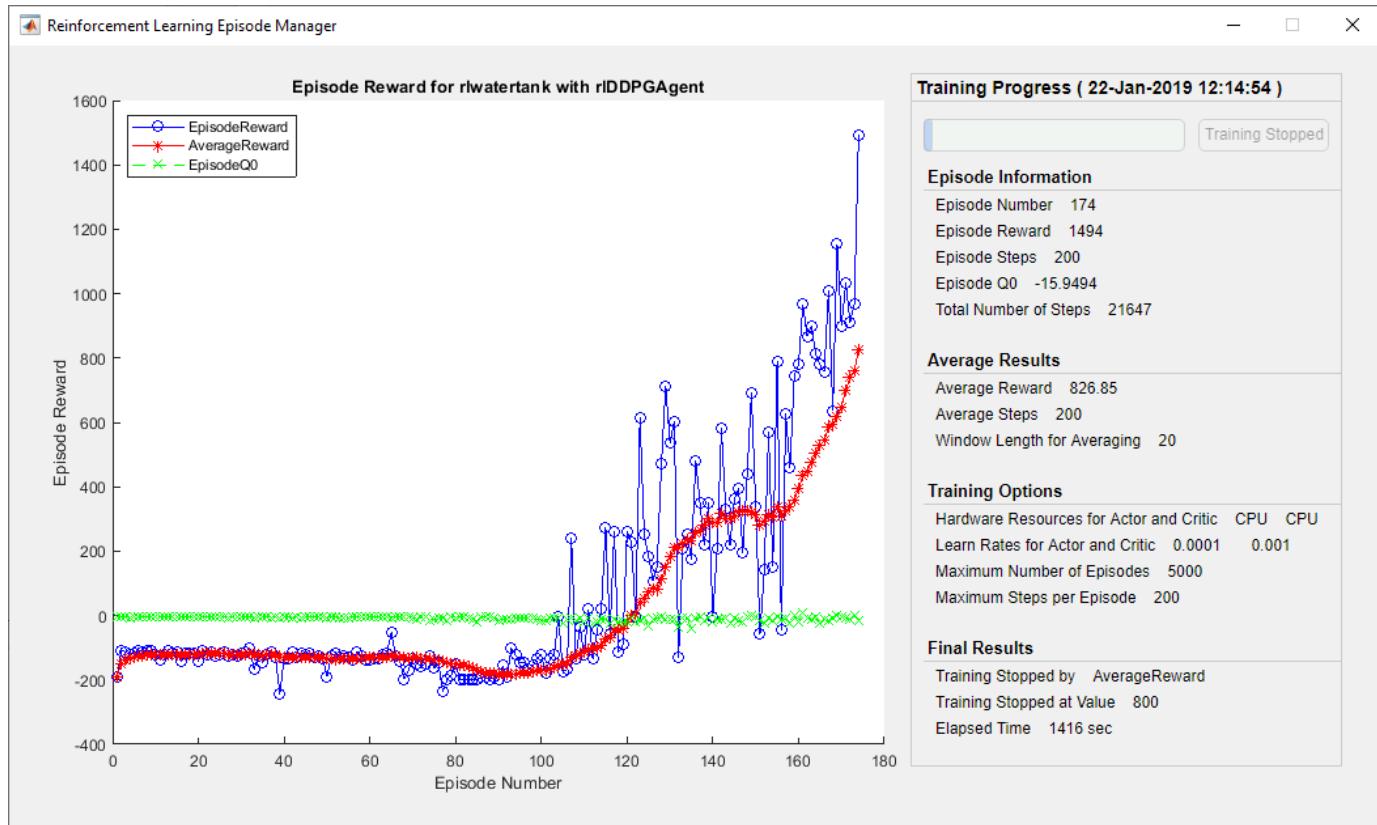
For more information, see `rlTrainingOptions`.

```
maxepisodes = 5000;
maxsteps = ceil(Tf/Ts);
trainOpts = rlTrainingOptions(...
    'MaxEpisodes', maxepisodes, ...
    'MaxStepsPerEpisode', maxsteps, ...
    'ScoreAveragingWindowLength', 20, ...
    'Verbose', false, ...
    'Plots', 'training-progress',...
    'StopTrainingCriteria', 'AverageReward',...
    'StopTrainingValue', 800);
```

Train the agent using the `train` function. Training is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

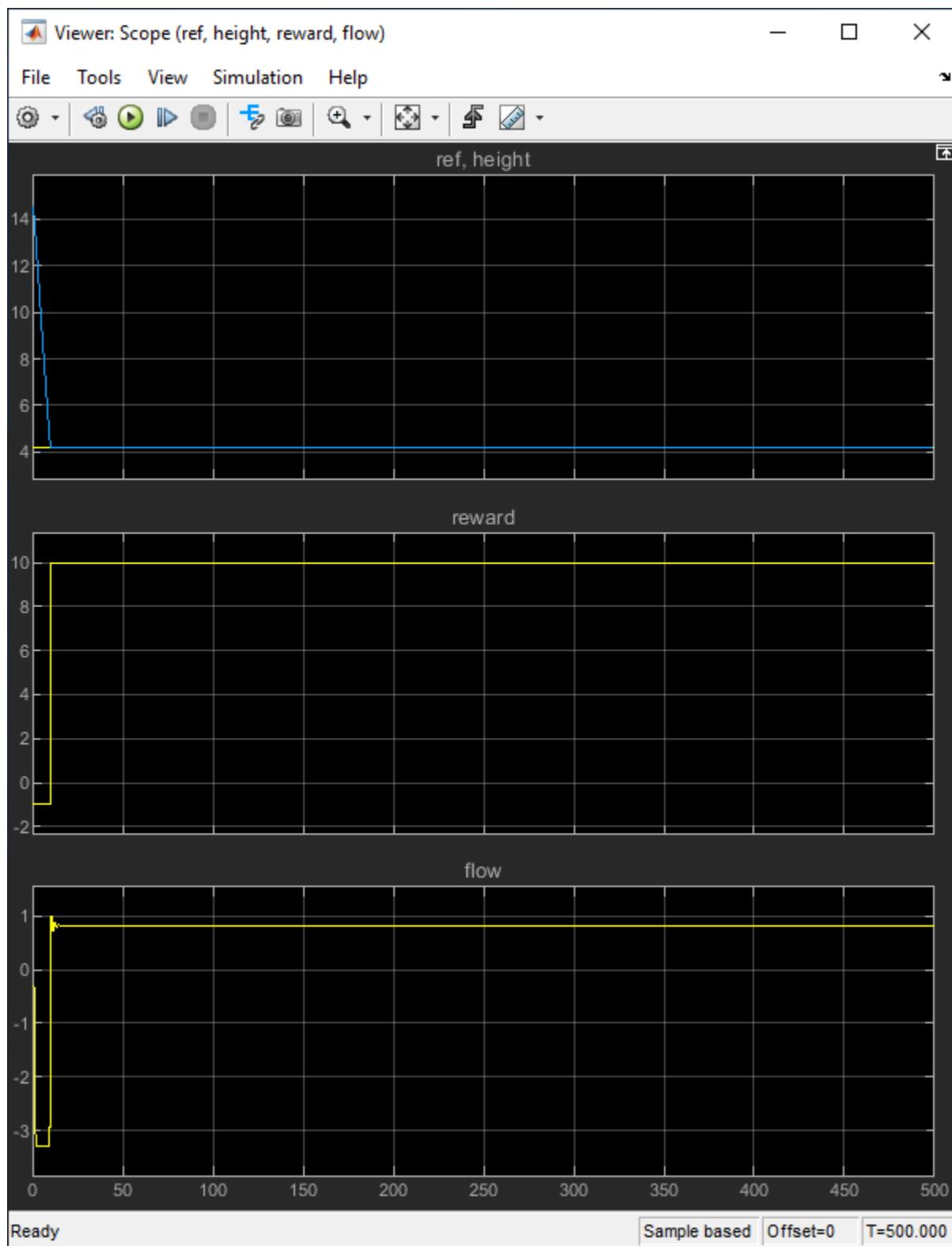
if doTraining
    % Train the agent.
    trainingStats = train(agent, env, trainOpts);
else
    % Load the pretrained agent for the example.
    load('WaterTankDDPG.mat', 'agent')
end
```



Validate Trained Agent

Validate the learned agent against the model by simulation.

```
simOpts = rlSimulationOptions('MaxSteps',maxsteps,'StopOnError','on');
experiences = sim(env,agent,simOpts);
```



Local Function

```
function in = localResetFcn(in)

% randomize reference signal
blk = sprintf('rlwatertank/Desired \nWater Level');
h = 3*randn + 10;
while h <= 0 || h >= 20
    h = 3*randn + 10;
end
in = setBlockParameter(in,blk,'Value',num2str(h));

% randomize initial height
h = 3*randn + 10;
while h <= 0 || h >= 20
    h = 3*randn + 10;
end
blk = 'rlwatertank/Water-Tank System/H';
in = setBlockParameter(in,blk,'InitialCondition',num2str(h));

end
```

See Also

[train](#)

More About

- “Train Reinforcement Learning Agents” on page 5-2
- “Create Simulink Environments for Reinforcement Learning” on page 2-5

Create Environments

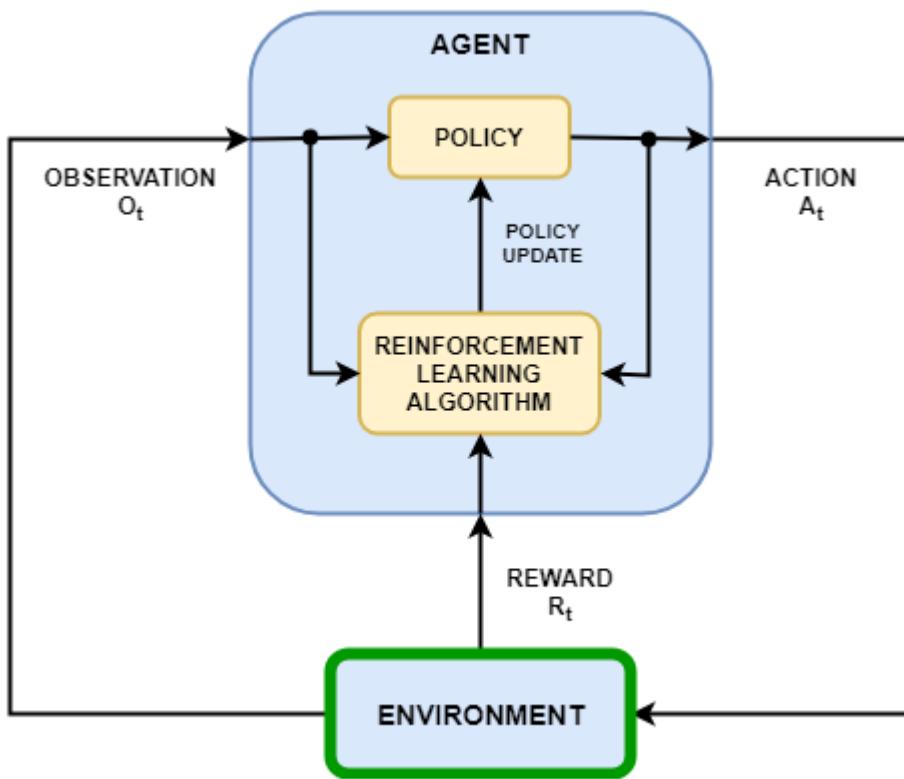
- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Define Reward Signals” on page 2-7
- “Load Predefined Grid World Environments” on page 2-9
- “Load Predefined Control System Environments” on page 2-15
- “Load Predefined Simulink Environments” on page 2-22
- “Create Custom Grid World Environments” on page 2-28
- “Create MATLAB Environment Using Custom Functions” on page 2-33
- “Create Custom MATLAB Environment from Template” on page 2-40
- “Water Tank Reinforcement Learning Environment Model” on page 2-46

Create MATLAB Environments for Reinforcement Learning

In a reinforcement learning scenario, where you are training an agent to complete a task, the environment models the external system (that is the world) with which the agent interacts. In control systems applications, this external system is often referred to as *the plant*.

As shown in the following figure, the environment:

- 1 Receives actions from the agent.
- 2 Returns observations in response to the actions.
- 3 Generates a reward measuring how well the action contributes to achieving the task.



Creating an environment model involves defining:

- Action and observation signals that the agent uses to interact with the environment.
- A reward signal that the agent uses to measure its success. For more information, see “Define Reward Signals” on page 2-7.
- The environment initial condition and its dynamic behavior.

Action and Observation Signals

When you create the environment object, you must specify the action and observation signals that the agent uses to interact with the environment. You can create both discrete and continuous action and observation spaces. For more information, see `rlNumericSpec` and `rlFiniteSetSpec`, respectively.

What signals you select as actions and observations depends on your application. For example, for control system applications, the integrals (and sometimes derivatives) of error signals are often useful observations. Also, for reference-tracking applications, having a time-varying reference signal as an observation is helpful.

When you define your observation signals, ensure that all the environment states (or their estimation) are included in the observation vector. This is a good practice because the agent is often a static function which lacks internal memory or state, and so it might not be able to successfully reconstruct the environment state internally.

For example, an image observation of a swinging pendulum has position information but does not have enough information, by itself, to determine the pendulum velocity. In this case, you can measure or estimate the pendulum velocity as an additional entry in the observation vector.

Predefined MATLAB Environments

The Reinforcement Learning Toolbox software provides some predefined MATLAB environments for which the actions, observations, rewards, and dynamics are already defined. You can use these environments to:

- Learn reinforcement learning concepts.
- Gain familiarity with Reinforcement Learning Toolbox software features.
- Test your own reinforcement learning agents.

For more information, see “Load Predefined Grid World Environments” on page 2-9 and “Load Predefined Control System Environments” on page 2-15.

Custom MATLAB Environments

You can create the following types of custom MATLAB environments for your own applications:

- Grid worlds with specified size, rewards, and obstacles
- Environments with dynamics specified using custom functions
- Environments specified by creating and modifying a template environment object

Once you create a custom environment object, you can train an agent in the same manner as in a predefined environment. For more information on training agents, see “Train Reinforcement Learning Agents” on page 5-2.

Custom Grid Worlds

You can create custom grid worlds of any size with your own custom reward, state transition, and obstacle configurations. To create a custom grid world environment:

- 1 Create a grid world model using the `createGridWorld` function. For example, create a grid world named `gw` with ten rows and nine columns.

```
gw = createGridWorld(10,9);
```

- 2 Configure the grid world by modifying the properties of the model. For example, specify the terminal state as the location [7,9]

```
gw.TerminalStates = "[7,9]";
```

- 3** A grid world needs to be included in a Markov decision process (MDP) environment. Create an MDP environment for this grid world, which the agent uses to interact with the grid world model.

```
env = rlMDPEnv(gw);
```

For more information on Custom Grid Worlds see “Create Custom Grid World Environments” on page 2-28.

Specify Custom Functions

For simple environments, you can define a custom environment object by creating an `rlFunctionEnv` object and specifying your own custom *reset* and *step* functions.

- At the beginning of each training episode, the agent calls the *reset* function to set the environment initial condition. For example, you can specify known initial state values or place the environment into a random initial state.
- The *step* function defines the dynamics of the environment, that is, how the state changes as a function of the current state and the agent action. At each training time step, the state of the model is updated using the *step* function.

For more information, see “Create MATLAB Environment Using Custom Functions” on page 2-33.

Create and Modify Template Environment

For more complex environments, you can define a custom environment by creating and modifying a template environment. To create a custom environment:

- 1 Create an environment template class using the `rlCreateEnvTemplate` function.
- 2 Modify the template environment, specifying environment properties, required environment functions, and optional environment functions.
- 3 Validate your custom environment using `validateEnvironment`.

For more information, see “Create Custom MATLAB Environment from Template” on page 2-40.

See Also

`rlCreateEnvTemplate` | `rlFunctionEnv` | `rlPredefinedEnv`

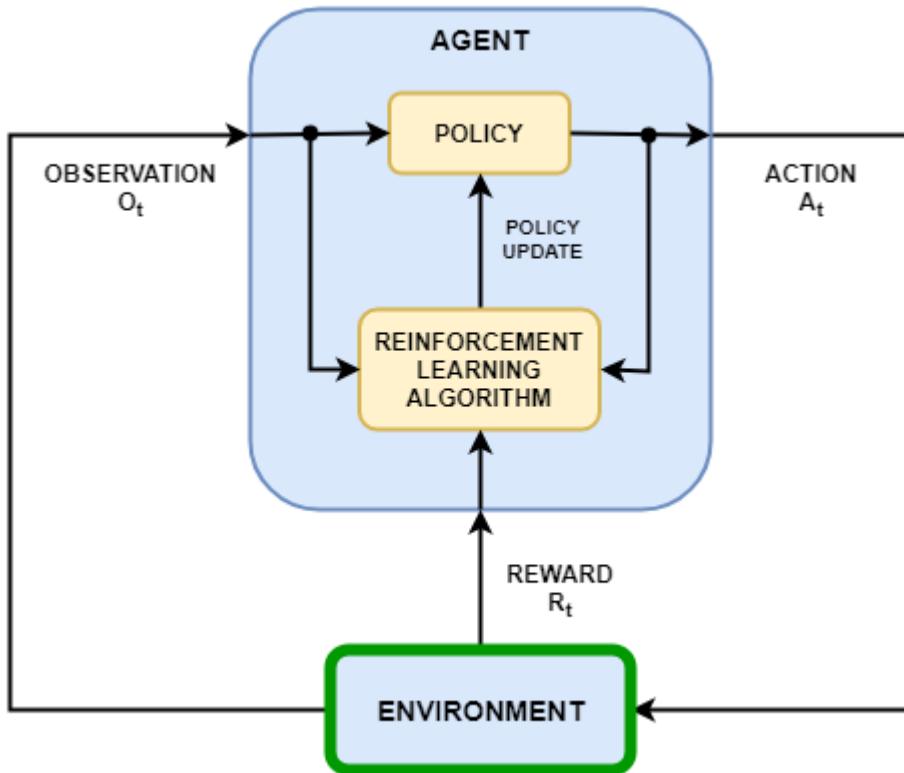
More About

- “What Is Reinforcement Learning?” on page 1-3
- “Create Simulink Environments for Reinforcement Learning” on page 2-5

Create Simulink Environments for Reinforcement Learning

In a reinforcement learning scenario, where you are training an agent to complete task, the environment models the dynamics with which the agent interacts. As shown in the following figure, the environment:

- 1 Receives actions from the agent.
- 2 Outputs observations in response to the actions.
- 3 Generates a reward measuring how well the action contributes to achieving the task.



Creating an environment model includes defining the following:

- Action and observation signals that the agent uses to interact with the environment.
- Reward signal that the agent uses to measure its success. For more information, see “Define Reward Signals” on page 2-7.
- Environment dynamic behavior.

Action and Observation Signals

When you create an environment object, you must specify the action and observation signals that the agent uses to interact with the environment. You can create both discrete and continuous action spaces. For more information, see `rlNumericSpec` and `rlFiniteSetSpec`, respectively.

What signals you select as actions and observations depends on your application. For example, for control system applications, the integrals (and sometimes derivatives) of error signals are often

useful observations. Also, for reference-tracking applications, having a time-varying reference signal as an observation is helpful.

When you define your observation signals, ensure that all the system states are observable through the observations. For example, an image observation of a swinging pendulum has position information but does not have enough information to determine the pendulum velocity. In this case, you can specify the pendulum velocity as a separate observation.

Predefined Simulink Environments

Reinforcement Learning Toolbox software provides predefined Simulink environments for which the actions, observations, rewards, and dynamics are already defined. You can use these environments to:

- Learn reinforcement learning concepts.
- Gain familiarity with Reinforcement Learning Toolbox software features.
- Test your own reinforcement learning agents.

For more information, see “Load Predefined Simulink Environments” on page 2-22.

Custom Simulink Environments

To specify your own custom reinforcement learning environment, create a Simulink model with an RL Agent block. In this model, connect the action, observation, and reward signals to the RL Agent block. For an example, see “Water Tank Reinforcement Learning Environment Model” on page 2-46.

For the action and observation signals, you must create specification objects using `rlNumericSpec` for continuous signals and `rlFiniteSetSpec` for discrete signals. For bus signals, create specifications using `bus2RLSpec`.

For the reward signal, construct a scalar signal in the model and connect this signal to the RL Agent block. For more information, see “Define Reward Signals” on page 2-7.

After configuring the Simulink model, create an environment object for the model using the `rlSimulinkEnv` function.

If you have a reference model with an appropriate action input port, observation output port, and scalar reward output port, you can automatically create a Simulink model that includes this reference model and an RL Agent block. For more information, see `createIntegratedEnv`. This function returns the environment object, action specifications, and observation specifications for the model.

Your environment can include third-party functionality. For more information, see “Integrate with Existing Simulation or Environment” (Simulink).

See Also

`createIntegratedEnv` | `rlPredefinedEnv` | `rlSimulinkEnv`

More About

- “What Is Reinforcement Learning?” on page 1-3
- “Create MATLAB Environments for Reinforcement Learning” on page 2-2

Define Reward Signals

To guide the learning process, reinforcement learning uses a scalar reward signal generated from the environment. This signal measures the performance of the agent with respect to the task goals. In other words, for a given observation (state), the reward measures the effectiveness of taking a particular action. During training, an agent updates its policy based on the rewards received for different state-action combinations. For more information on the different types of agents and how they use the reward signal during training, see “Reinforcement Learning Agents” on page 4-2.

In general, you provide a positive reward to encourage certain agent actions and a negative reward (penalty) to discourage other actions. A well-designed reward signal guides the agent to maximize the expectation of the long-term reward. What constitutes a well-designed reward depends on your application and the agent goals.

For example, when an agent must perform a task for as long as possible, a common strategy is to provide a small positive reward for each time step that the agent successfully performs the task and a large penalty when the agent fails. This approach encourages longer training episodes while heavily discouraging episodes that fail. For an example that uses this approach, see “Train DQN Agent to Balance Cart-Pole System” on page 5-8.

If your reward function incorporates multiple signals, such as position, velocity, and control effort, you must consider the relative sizes of the signals and scale their contributions to the reward signal accordingly.

You can specify either continuous or discrete reward signals. In either case, you must provide a reward signal that provides rich information when the action and observation signals change.

Continuous Rewards

A continuous reward function varies continuously with changes in the environment observations and actions. In general, continuous reward signals improve convergence during training and can lead to simpler network structures.

An example of a continuous reward is the quadratic regulator (QR) cost function, where the long-term reward can be expressed as

$$J_i = - \left(s_\tau^T Q_\tau s_\tau + \sum_{j=i}^{\tau} s_j^T Q_j s_j + a_j^T R_j a_j + 2s_j^T N_j a_j \right)$$

Here, Q_τ , Q , R , and N are the weight matrices. Q_τ is the terminal weight matrix, applied only at the end of the episode. Also, s is the observation vector, a is the action vector, and τ is the terminal iteration of the episode. The instantaneous reward for this cost function is

$$r_i = s_i^T Q_i s_i + a_i^T R_i a_i + 2s_i^T N_i a_i$$

This QR reward structure encourages driving s to zero with minimal action effort. A QR-based reward structure is a good reward to choose for regulation or stationary point problems, such as pendulum swing-up or regulating the position of the double integrator. For training examples that use a QR reward, see “Train DQN Agent to Swing Up and Balance Pendulum” on page 5-37 and “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Smooth continuous rewards, such as the QR regulator, are good for fine-tuning parameters and can provide policies similar to optimal controllers (LQR/MPC).

Discrete Rewards

A discrete reward function varies discontinuously with changes in the environment observations or actions. These types of reward signals can make convergence slower and can require more complex network structures. Discrete rewards are usually implemented as *events* that occur in the environment—for example, when an agent receives a positive reward if it exceeds some target value or a penalty when it violates some performance constraint.

While discrete rewards can slow down convergence, they can also guide the agent toward better reward regions in the state space of the environment. For example, a region-based reward, such as a fixed reward when the agent is near a target location, can emulate final-state constraints. Also, a region-based penalty can encourage an agent to avoid certain areas of the state space.

Mixed Rewards

In many cases, providing a mixed reward signal that has a combination of continuous and discrete reward components is beneficial. The discrete reward signal can be used to drive the system away from bad states, and the continuous reward signal can improve convergence by providing a smooth reward near target states. For example, in “Train DDPG Agent to Control Flying Robot” on page 5-89, the reward function has three components: r_1 , r_2 , and r_3 .

$$\begin{aligned}r_1 &= 10((x_t^2 + y_t^2 + \theta_t^2) < 0.5) \\r_2 &= -100(|x_t| \geq 20 \text{ } || |y_t| \geq 20) \\r_3 &= -(0.2(R_{t-1} + L_{t-1})^2 + 0.3(R_{t-1} - L_{t-1})^2 + 0.03x_t^2 + 0.03y_t^2 + 0.02\theta_t^2) \\r &= r_1 + r_2 + r_3\end{aligned}$$

Here:

- r_1 is a region-based continuous reward that applies only near the target location of the robot.
- r_2 is a discrete signal that provides a large penalty when the robot moves far from the target location.
- r_3 is a continuous QR penalty that applies for all robot states.

See Also

More About

- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Create Simulink Environments for Reinforcement Learning” on page 2-5

Load Predefined Grid World Environments

Reinforcement Learning Toolbox software provides several predefined grid world environments for which the actions, observations, rewards, and dynamics are already defined. You can use these environments to:

- Learn reinforcement learning concepts.
- Gain familiarity with Reinforcement Learning Toolbox software features.
- Test your own reinforcement learning agents.

You can load the following predefined MATLAB grid world environments using the `rlPredefinedEnv` function.

Environment	Agent Task
Basic grid world	Move from a starting location to a target location on a two-dimensional grid by selecting moves from the discrete action space $\{N, S, E, W\}$.
Waterfall grid world	Move from a starting location to a target location on a larger two-dimensional grid with unknown deterministic or stochastic dynamics.

For more information on the properties of grid world environments, see “Create Custom Grid World Environments” on page 2-28.

You can also load predefined MATLAB control system environments. For more information, see “Load Predefined Control System Environments” on page 2-15.

Basic Grid World

The basic grid world environment is a two-dimensional 5-by-5 grid with a starting location, terminal location, and obstacles. The environment also contains a special jump from state [2,4] to state [4,4]. The goal of the agent is to move from the starting location to the terminal location while avoiding obstacles and maximizing the total reward.

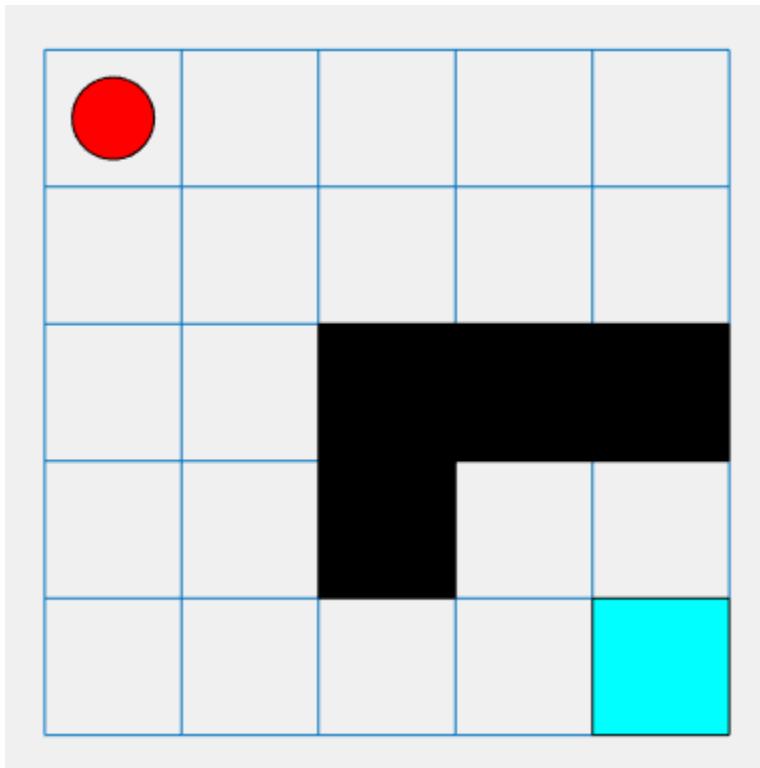
To create a basic grid world environment, use the `rlPredefinedEnv` function. This function creates an `rlMDPEnv` object representing the grid world.

```
env = rlPredefinedEnv('BasicGridWorld');
```

You can visualize the grid world environment using the `plot` function.

- Agent location is a red circle. By default, the agent starts in state [1,1].
- Terminal location is a blue square.
- Obstacles are black squares.

```
plot(env)
```



Actions

The agent can move in one of four possible directions (north, south, east, or west).

Rewards

The agent receives the following rewards or penalties:

- +10 reward for reaching the terminal state at [5,5]
- +5 reward for jumping from state [2,4] to state [4,4]
- -1 penalty for every other action

Deterministic Waterfall Grid Worlds

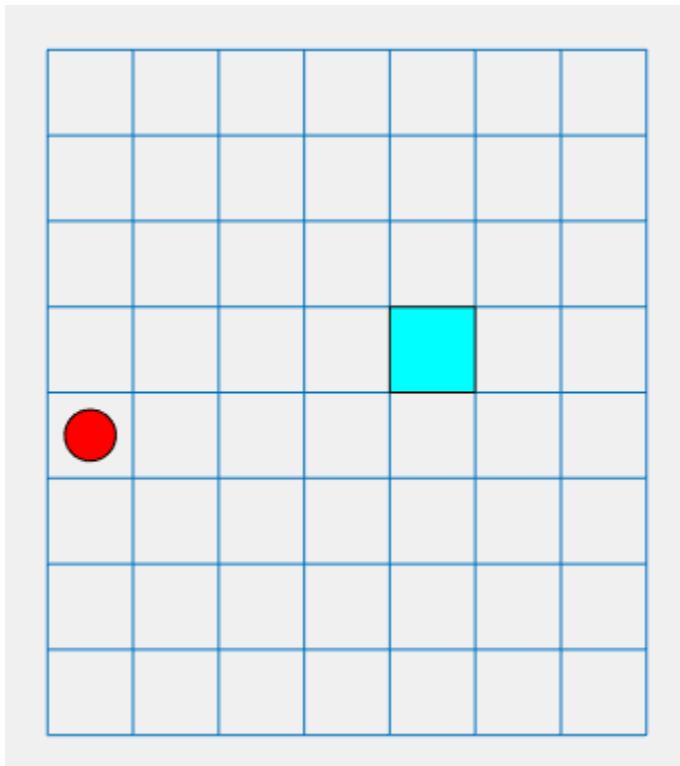
The deterministic waterfall grid world environment is a two-dimensional 8-by-7 grid with a starting location and terminal location. The environment includes a waterfall that pushes the agent toward the bottom of the grid. The goal of the agent is to move from the starting location to the terminal location while maximizing the total reward.

To create a deterministic waterfall grid world, use the `rlPredefinedEnv` function. This function creates an `rlMDPEnv` object representing the grid world.

```
env = rlPredefinedEnv('WaterFallGridWorld-Deterministic');
```

As with the basic grid world, you can visualize the environment, where the agent is a red circle and the terminal location is a blue square.

```
plot(env)
```



Actions

The agent can move in one of four possible directions (north, south, east, or west).

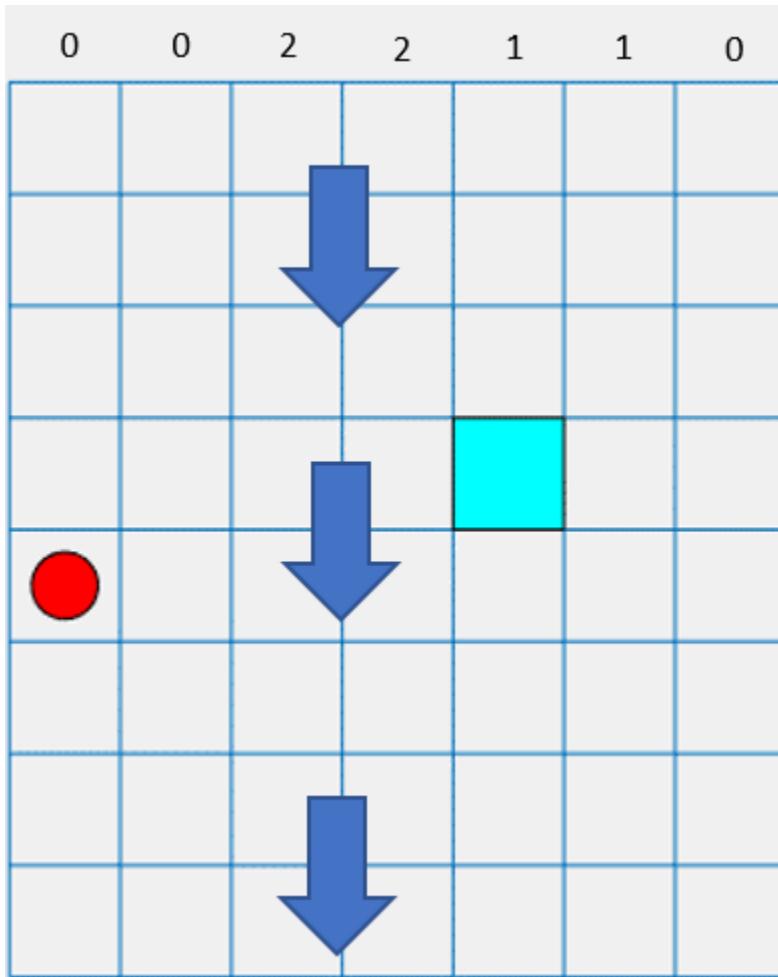
Rewards

The agent receives the following rewards or penalties:

- +10 reward for reaching the terminal state at [4,5]
- -1 penalty for every other action

Waterfall Dynamics

In this environment, a waterfall pushes the agent toward the bottom of the grid.



The intensity of the waterfall varies between the columns, as shown at the top of the preceding figure. When the agent moves into a column with a nonzero intensity, the waterfall pushes it downward by the indicated number of squares. For example, if the agent goes east from state [5,2], it reaches state [7,3].

Stochastic Waterfall Grid Worlds

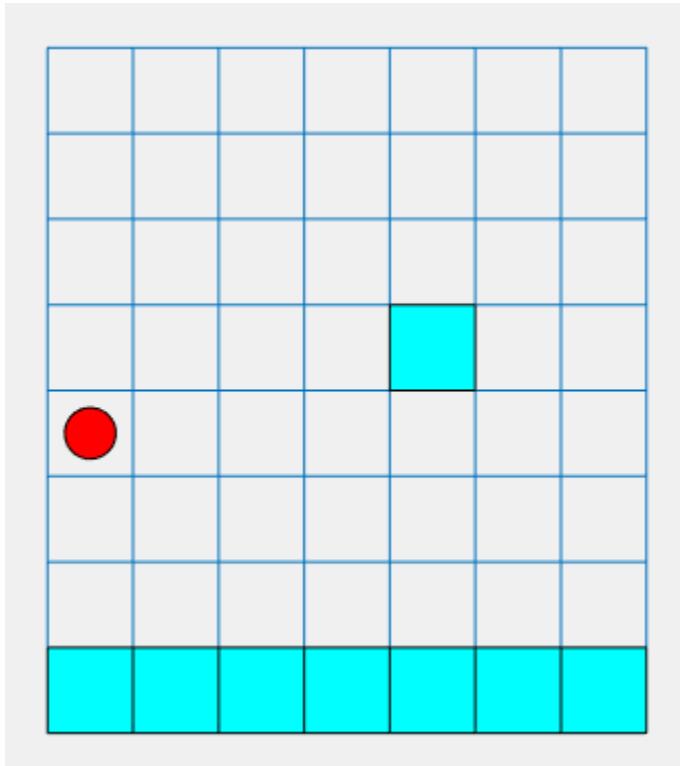
The stochastic waterfall grid world environment is a two-dimensional 8-by-7 grid with a starting location and terminal locations. The environment includes a waterfall that pushes the agent towards the bottom of the grid with a stochastic intensity. The goal of the agent is to move from the starting location to the target terminal location while avoiding the penalty terminal states along the bottom of the grid and maximizing the total reward.

To create a stochastic waterfall grid world, use the `rlPredefinedEnv` function. This function creates an `rlMDPEnv` object representing the grid world.

```
env = rlPredefinedEnv('WaterFallGridWorld-Stochastic');
```

As with the basic grid world, you can visualize the environment, where the agent is a red circle and the terminal location is a blue square.

```
plot(env)
```



Actions

The agent can move in one of four possible directions (north, south, east, or west).

Rewards

The agent receives the following rewards or penalties:

- +10 reward for reaching the terminal state at [4,5]
- - 10 penalty for reaching any terminal state in the bottom row of the grid
- - 1 penalty for every other action

Waterfall Dynamics

In this environment, a waterfall pushes the agent towards the bottom of the grid with a stochastic intensity. The baseline intensity matches the intensity of the deterministic waterfall environment. However, in the stochastic waterfall case, the agent has an equal chance of experiencing the indicated intensity, one level above that intensity, or one level below that intensity. For example, if the agent goes east from state [5,2], it has an equal chance of reaching state [6,3], [7,3], or [8,3].

See Also

[rlMDPEnv](#) | [rlPredefinedEnv](#) | [train](#)

More About

- “Create MATLAB Environments for Reinforcement Learning” on page 2-2

- “Load Predefined Control System Environments” on page 2-15
- “Train Reinforcement Learning Agents” on page 5-2

Load Predefined Control System Environments

Reinforcement Learning Toolbox software provides several predefined control system environments for which the actions, observations, rewards, and dynamics are already defined. You can use these environments to:

- Learn reinforcement learning concepts.
- Gain familiarity with Reinforcement Learning Toolbox software features.
- Test your own reinforcement learning agents.

You can load the following predefined MATLAB control system environments using the `rlPredefinedEnv` function.

Environment	Agent Task
Cart-pole	Balance a pole on a moving cart by applying forces to the cart using either a discrete or continuous action space.
Double integrator	Control a second-order dynamic system using either a discrete or continuous action space.
Simple pendulum with image observation	Swing up and balance a simple pendulum using either a discrete or continuous action space.

You can also load predefined MATLAB grid world environments. For more information, see “Load Predefined Grid World Environments” on page 2-9.

Cart-Pole Environments

The goal of the agent in the predefined cart-pole environments is to balance a pole on a moving cart by applying horizontal forces to the cart. The pole is considered successfully balanced if both of the following conditions are satisfied:

- The pole angle remains within a given threshold of the vertical position, where the vertical position is zero radians.
- The magnitude of the cart position remains below a given threshold.

There are two cart-pole environment variants, which differ by the agent action space.

- Discrete — Agent can apply a force of either F_{max} or $-F_{max}$ to the cart, where F_{max} is the `MaxForce` property of the environment.
- Continuous — Agent can apply any force within the range $[-F_{max}, F_{max}]$.

To create a cart-pole environment, use the `rlPredefinedEnv` function.

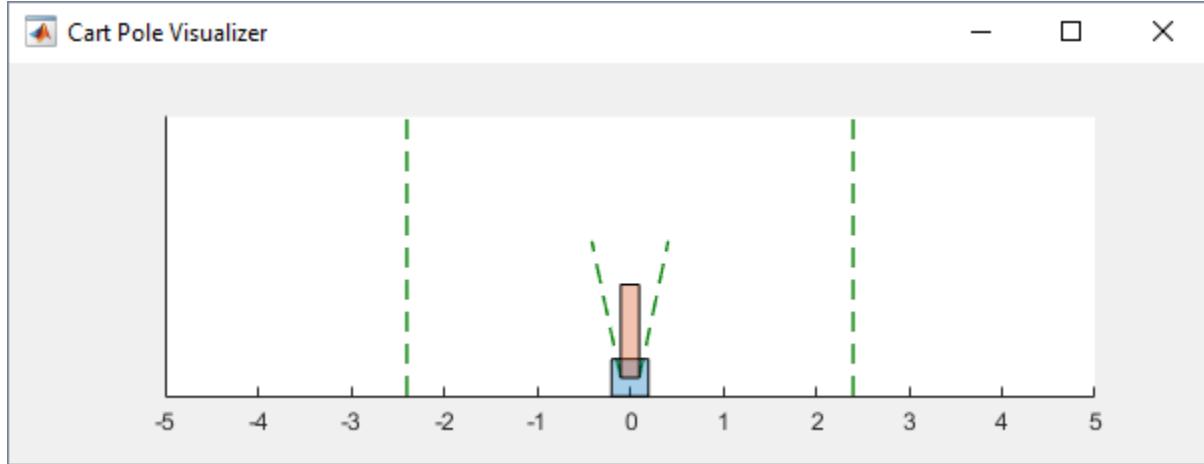
- Discrete action space

```
env = rlPredefinedEnv('CartPole-Discrete');
```
- Continuous action space

```
env = rlPredefinedEnv('CartPole-Continuous');
```

You can visualize the cart-pole environment using the `plot` function. The plot displays the cart as a blue square and the pole as a red rectangle.

```
plot(env)
```



To visualize the environment during training, call `plot` before training and keep the visualization figure open.

For examples showing how to train agents in cart-pole environments, see the following:

- “Train DQN Agent to Balance Cart-Pole System” on page 5-8
- “Train PG Agent to Balance Cart-Pole System” on page 5-14
- “Train AC Agent to Balance Cart-Pole System” on page 5-19

Environment Properties

Property	Description	Default
Gravity	Acceleration due to gravity in meters per second	9.8
MassCart	Mass of the cart in kilograms	1
MassPole	Mass of the pole in kilograms	0.1
Length	Half the length of the pole in meters	0.5
MaxForce	Maximum horizontal force magnitude in newtons	10
Ts	Sample time in seconds	0.02
ThetaThresholdRadians	Pole angle threshold in radians	0.2094
XThreshold	Cart position threshold in meters	2.4
RewardForNotFalling	Reward for each time step the pole is balanced	1
PenaltyForFalling	Reward penalty for failing to balance the pole	Discrete — -5 Continuous — -50

Property	Description	Default
State	<p>Environment state, specified as a column vector with the following state variables:</p> <ul style="list-style-type: none"> • Cart position • Derivative of cart position • Pole angle • Derivative of pole angle 	[0 0 0 0]'

Actions

In the cart-pole environments, the agent interacts with the environment using a single action signal, the horizontal force applied to the cart. The environment contains a specification object for this action signal. For the environment with a:

- Discrete action space, the specification is an `rlFiniteSetSpec` object.
- Continuous action space, the specification is an `rlNumericSpec` object.

For more information on obtaining action specifications from an environment, see `getActionInfo`.

Observations

In the cart-pole system, the agent can observe all the environment state variables in `env.State`. For each state variable, the environment contains an `rlNumericSpec` observation specification. All the states are continuous and unbounded.

For more information on obtaining observation specifications from an environment, see `getObservationInfo`.

Reward

The reward signal for this environment consists of two components.

- A positive reward for each time step that the pole is balanced, that is, the cart and pole both remain within their specified threshold ranges. This reward accumulates over the entire training episode. To control the size of this reward, use the `RewardForNotFalling` property of the environment.
- A one-time negative penalty if either the pole or cart moves outside of their threshold range. At this point, the training episode stops. To control the size of this penalty, use the `PenaltyForFalling` property of the environment.

Double Integrator Environments

The goal of the agent in the predefined double integrator environments is to control the position of a mass in a second-order system by applying a force input. Specifically, the second-order system is a double integrator with a gain.

Training episodes for these environments end when either of the following events occurs:

- The mass moves beyond a given threshold from the origin.
- The norm of the state vector is less than a given threshold.

There are two double integrator environment variants, which differ by the agent action space.

- Discrete — Agent can apply a force of either F_{max} or $-F_{max}$ to the cart, where F_{max} is the `MaxForce` property of the environment.
- Continuous — Agent can apply any force within the range $[-F_{max}, F_{max}]$.

To create a double integrator environment, use the `rlPredefinedEnv` function.

- Discrete action space

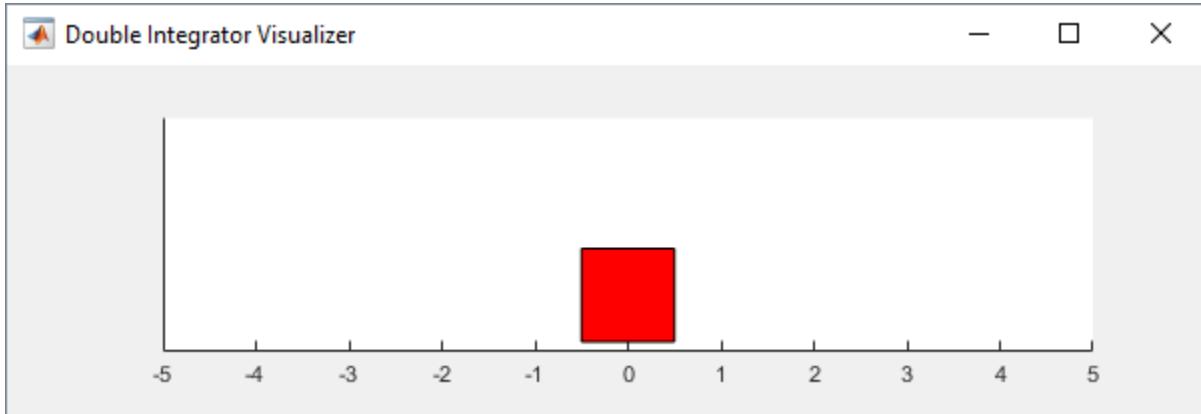
```
env = rlPredefinedEnv('DoubleIntegrator-Discrete');
```

- Continuous action space

```
env = rlPredefinedEnv('DoubleIntegrator-Continuous');
```

You can visualize the double integrator environment using the `plot` function. The plot displays the mass as a red rectangle.

```
plot(env)
```



To visualize the environment during training, call `plot` before training and keep the visualization figure open.

For examples showing how to train agents in double integrator environments, see the following:

- “Train DDPG Agent to Control Double Integrator System” on page 5-31
- “Train PG Agent with Baseline to Control Double Integrator System” on page 5-25

Environment Properties

Property	Description	Default
<code>Gain</code>	Gain for the double integrator	1
<code>Ts</code>	Sample time in seconds	0.1
<code>MaxDistance</code>	Distance magnitude threshold in meters	5
<code>GoalThreshold</code>	State norm threshold	0.01
<code>Q</code>	Weight matrix for observation component of reward signal	[10 0; 0 1]

Property	Description	Default
R	Weight matrix for action component of reward signal	0.01
MaxForce	Maximum input force in newtons	Discrete: 2 Continuous: Inf
State	Environment state, specified as a column vector with the following state variables: <ul style="list-style-type: none">• Mass position• Derivative of mass position	[0 0]'

Actions

In the double integrator environments, the agent interacts with the environment using a single action signal, the force applied to the mass. The environment contains a specification object for this action signal. For the environment with a:

- Discrete action space, the specification is an `rlFiniteSetSpec` object.
- Continuous action space, the specification is an `rlNumericSpec` object.

For more information on obtaining action specifications from an environment, see `getActionInfo`.

Observations

In the double integrator system, the agent can observe both of the environment state variables in `env.State`. For each state variable, the environment contains an `rlNumericSpec` observation specification. Both states are continuous and unbounded.

For more information on obtaining observation specifications from an environment, see `getObservationInfo`.

Reward

The reward signal for this environment is the discrete-time equivalent of the following continuous-time reward, which is analogous to the cost function of an LQR controller.

$$\text{reward} = - \int (x'Qx + u'Ru) dt$$

Here:

- Q and R are environment properties.
- x is the environment state vector.
- u is the input force.

This reward is the episodic reward, that is, the cumulative reward across the entire training episode.

Simple Pendulum Environments with Image Observation

This environment is a simple frictionless pendulum that is initially hangs in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

There are two simple pendulum environment variants, which differ by the agent action space.

- Discrete — Agent can apply a torque of -2, -1, 0, 1, or 2 to the pendulum.
- Continuous — Agent can apply any torque within the range [-2,2].

To create a simple pendulum environment, use the `rlPredefinedEnv` function.

- Discrete action space

```
env = rlPredefinedEnv('SimplePendulumWithImage-Discrete');
```

- Continuous action space

```
env = rlPredefinedEnv('SimplePendulumWithImage-Continuous');
```

For examples showing how to train an agent in this environment, see the following:

- “Train DDPG Agent to Swing Up and Balance Pendulum with Image Observation” on page 5-64
- “Create Agent Using Deep Network Designer and Train Using Image Observations” on page 5-72

Environment Properties

Property	Description	Default
Mass	Pendulum mass	1
RodLength	Pendulum length	1
RodInertia	Pendulum moment of inertia	0
Gravity	Acceleration due to gravity in meters per second	9.81
DampingRatio	Damping on pendulum motion	0
MaximumTorque	Maximum input torque in newtons	2
Ts	Sample time in seconds	0.05
State	Environment state, specified as a column vector with the following state variables: <ul style="list-style-type: none"> • Pendulum angle • Pendulum angular velocity 	[0 0]'
Q	Weight matrix for observation component of reward signal	[1 0; 0 0.1]
R	Weight matrix for action component of reward signal	1e-3

Actions

In the simple pendulum environments, the agent interacts with the environment using a single action signal, the torque applied at the base of the pendulum. The environment contains a specification object for this action signal. For the environment with a:

- Discrete action space, the specification is an `rlFiniteSetSpec` object.
- Continuous action space, the specification is an `rlNumericSpec` object.

For more information on obtaining action specifications from an environment, see `getActionInfo`.

Observations

In the simple pendulum environment, the agent receives the following observation signals:

- 50-by-50 grayscale image of the pendulum position
- Derivative of the pendulum angle

For each observation signal, the environment contains an `rlNumericSpec` observation specification. All the observations are continuous and unbounded.

For more information on obtaining observation specifications from an environment, see [getObservationInfo](#).

Reward

The reward signal for this environment is

$$r_t = -(\theta_t^2 + 0.1 * \dot{\theta}_t^2 + 0.001 * u_{t-1}^2)$$

Here:

- θ_t is the pendulum angle of displacement from the upright position.
- $\dot{\theta}_t$ is the derivative of the pendulum angle.
- u_{t-1} is the control effort from the previous time step.

See Also

[rlPredefinedEnv](#) | [train](#)

More About

- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Load Predefined Grid World Environments” on page 2-9
- “Train Reinforcement Learning Agents” on page 5-2

Load Predefined Simulink Environments

Reinforcement Learning Toolbox software provides predefined Simulink environments for which the actions, observations, rewards, and dynamics are already defined. You can use these environments to:

- Learn reinforcement learning concepts.
- Gain familiarity with Reinforcement Learning Toolbox software features
- Test your own reinforcement learning agents.

You can load the following predefined Simulink environments using the `rlPredefinedEnv` function.

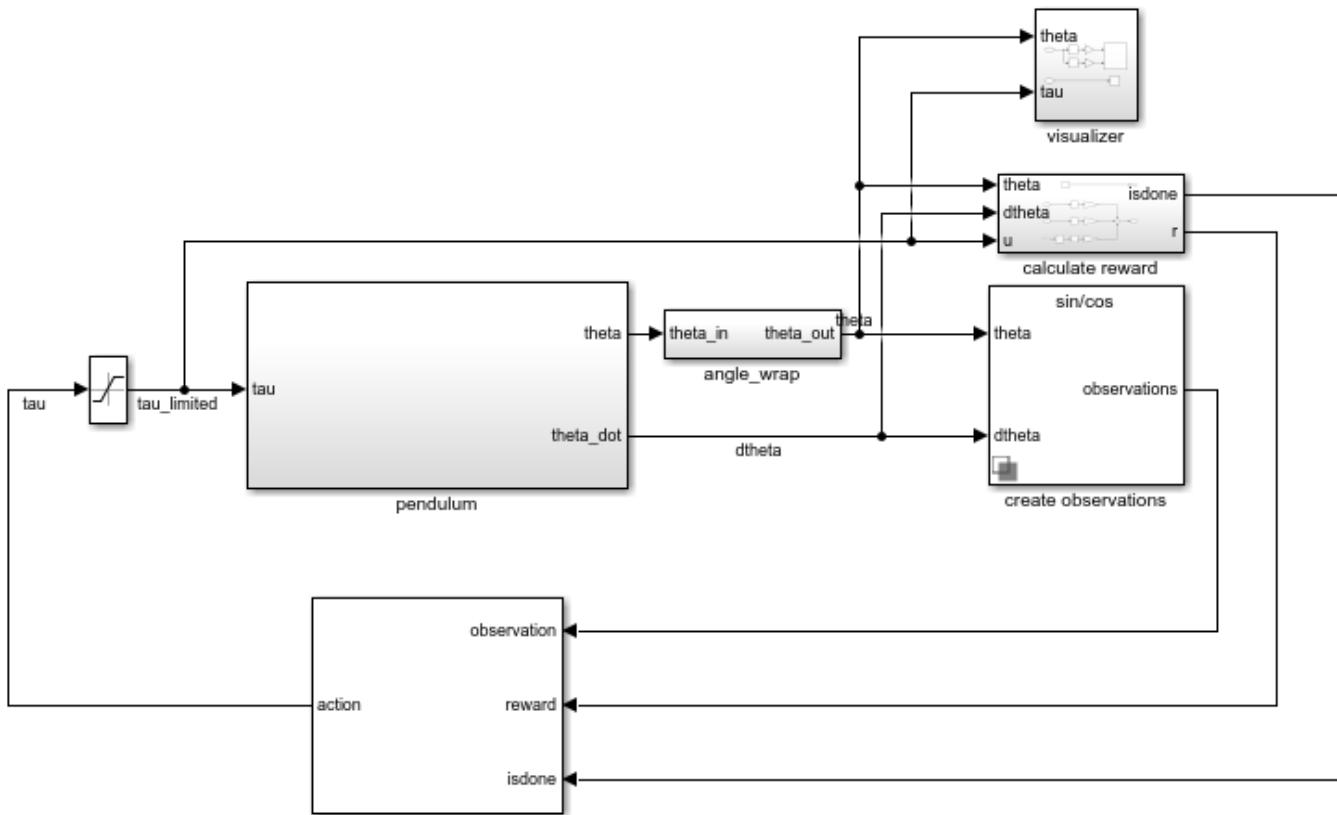
Environment	Agent Task
Simple pendulum Simulink model	Swing up and balance a simple pendulum using either a discrete or continuous action space.
Cart-pole Simscape™ model	Balance a pole on a moving cart by applying forces to the cart using either a discrete or continuous action space.

For predefined Simulink environments, the environment dynamics, observations, and reward signal are defined in a corresponding Simulink model. The `rlPredefinedEnv` function creates a `SimulinkEnvWithAgent` object that the `train` function uses to interact with the Simulink model.

Simple Pendulum Simulink Model

This environment is a simple frictionless pendulum that initially hangs in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort. The model for this environment is defined in the `rlSimplePendulumModel` Simulink model.

```
open_system('rlSimplePendulumModel')
```



There are two simple pendulum environment variants, which differ by the agent action space.

- Discrete — Agent can apply a torque of either T_{max} , 0 , or $-T_{max}$ to the pendulum, where T_{max} is the `max_tau` variable in the model workspace.
- Continuous — Agent can apply any torque within the range $[-T_{max}, T_{max}]$.

To create a simple pendulum environment, use the `rlPredefinedEnv` function.

- Discrete action space

```
env = rlPredefinedEnv('SimplePendulumModel-Discrete');
```

- Continuous action space

```
env = rlPredefinedEnv('SimplePendulumModel-Continuous');
```

For examples that train agents in the simple pendulum environment, see:

- “Train DQN Agent to Swing Up and Balance Pendulum” on page 5-37
- “Train DDPG Agent to Swing Up and Balance Pendulum” on page 5-43

Actions

In the simple pendulum environments, the agent interacts with the environment using a single action signal, the torque applied at the base of the pendulum. The environment contains a specification object for this action signal. For the environment with a:

- Discrete action space, the specification is an `rlFiniteSetSpec` object.
- Continuous action space, the specification is an `rlNumericSpec` object.

For more information on obtaining action specifications from an environment, see `getActionInfo`.

Observations

In the simple pendulum environment, the agent receives the following three observation signals, which are constructed within the create observations subsystem.

- Sine of the pendulum angle
- Cosine of the pendulum angle
- Derivative of the pendulum angle

For each observation signal, the environment contains an `rlNumericSpec` observation specification. All the observations are continuous and unbounded.

For more information on obtaining observation specifications from an environment, see `getObservationInfo`.

Reward

The reward signal for this environment, which is constructed in the calculate reward subsystem, is

$$r_t = - \left(\theta_t^2 + 0.1 * \dot{\theta}_t^2 + 0.001 * u_{t-1}^2 \right)$$

Here:

- θ_t is the pendulum angle of displacement from the upright position.
- $\dot{\theta}_t$ is the derivative of the pendulum angle.
- u_{t-1} is the control effort from the previous time step.

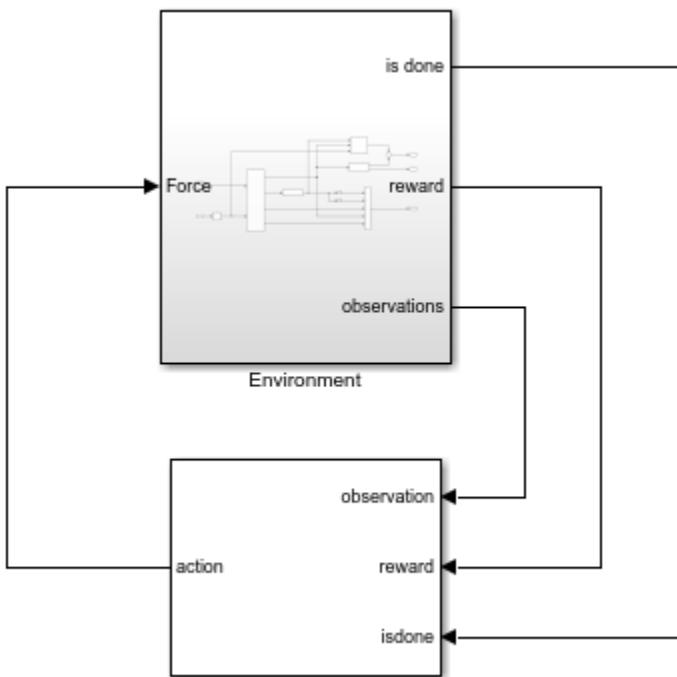
Cart-Pole Simscape Model

The goal of the agent in the predefined cart-pole environments is to balance a pole on a moving cart by applying horizontal forces to the cart. The pole is considered successfully balanced if both of the following conditions are satisfied:

- The pole angle remains within a given threshold of the vertical position, where the vertical position is zero radians.
- The magnitude of the cart position remains below a given threshold.

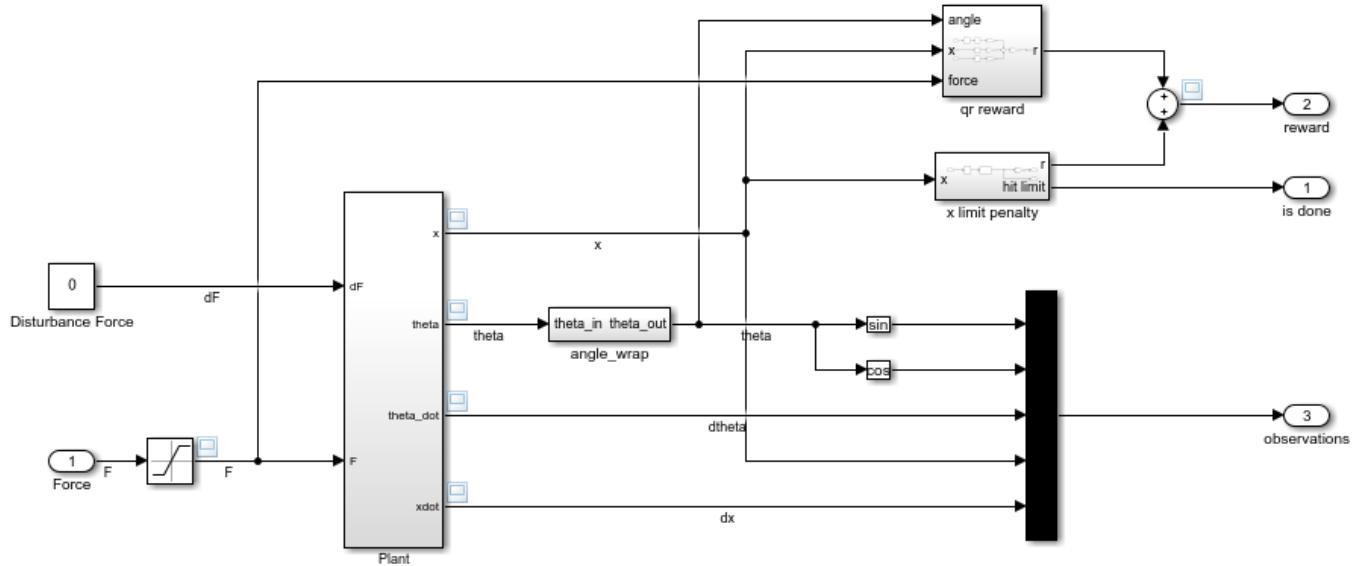
The model for this environment is defined in the `rlCartPoleSimscapeModel` Simulink model. The dynamics of this model are defined using Simscape Multibody™.

```
open_system('rlCartPoleSimscapeModel')
```



In the Environment subsystem, the model dynamics are defined using Simscape components and the reward and observation are constructed using Simulink blocks.

```
open_system('rlCartPoleSimscapeModel/Environment')
```



There are two cart-pole environment variants, which differ by the agent action space.

- Discrete — Agent can apply a force of 15, 0, or -15 to the cart.
- Continuous — Agent can apply any force within the range [-15,15].

To create a cart-pole environment, use the `rlPredefinedEnv` function.

- Discrete action space

```
env = rlPredefinedEnv('CartPoleSimscapeModel-Discrete');
```

- Continuous action space

```
env = rlPredefinedEnv('CartPoleSimscapeModel-Continuous');
```

For an example that trains an agent in this cart-pole environment, see “Train DDPG Agent to Swing Up and Balance Cart-Pole System” on page 5-50.

Actions

In the cart-pole environments, the agent interacts with the environment using a single action signal, the force applied to the cart. The environment contains a specification object for this action signal. For the environment with a:

- Discrete action space, the specification is an `rlFiniteSetSpec` object.
- Continuous action space, the specification is an `rlNumericSpec` object.

For more information on obtaining action specifications from an environment, see `getActionInfo`.

Observations

In the cart-pole environment, the agent receives the following five observation signals.

- Sine of the pole angle
- Cosine of the pole angle
- Derivative of the pendulum angle
- Cart position
- Derivative of cart position

For each observation signal, the environment contains an `rlNumericSpec` observation specification. All the observations are continuous and unbounded.

For more information on obtaining observation specifications from an environment, see `getObservationInfo`.

Reward

The reward signal for this environment is the sum of three components ($r = r_{qr} + r_n + r_p$):

- A quadratic regulator control reward, constructed in the `Environment/qr` reward subsystem.

$$r_{qr} = - (0.1 * x^2 + 0.5 * \theta^2 + 0.005 * u_{t-1}^2)$$

- An additional reward for when the pole is near the upright position, constructed in the `Environment/near upright` reward subsystem.

$$r_n = 10 * (|\theta| < 0.175)$$

- A cart limit penalty, constructed in the `Environment/x limit` penalty subsystem. This subsystem generates a negative reward when the magnitude of the cart position exceeds a given threshold.

$$r_p = -100 * (|x| \geq 3.5)$$

Here:

- x is the cart position.
- θ is the pole angle of displacement from the upright position.
- u_{t-1} is the control effort from the previous time step.

See Also

Blocks

RL Agent

Functions

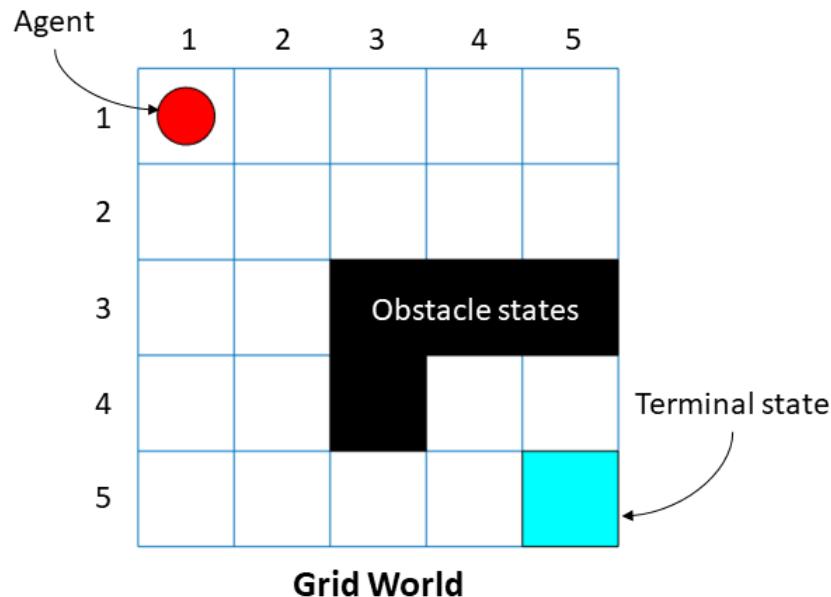
`rlPredefinedEnv` | `train`

More About

- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Train Reinforcement Learning Agents” on page 5-2

Create Custom Grid World Environments

A grid world is a two-dimensional, cell-based environment where the agent starts from one cell and moves toward the terminal cell while collecting as much reward as possible. Grid world environments are useful for applying reinforcement learning algorithms to discover optimal paths and policies for agents on the grid to arrive at the terminal goal in the fewest moves.



Reinforcement Learning Toolbox lets you create custom MATLAB grid world environments for your own applications. To create a custom grid world environment:

- 1 Create the grid world model.
- 2 Configure the grid world model.
- 3 Use the grid world model to create your own grid world environment.

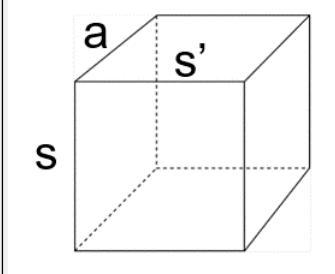
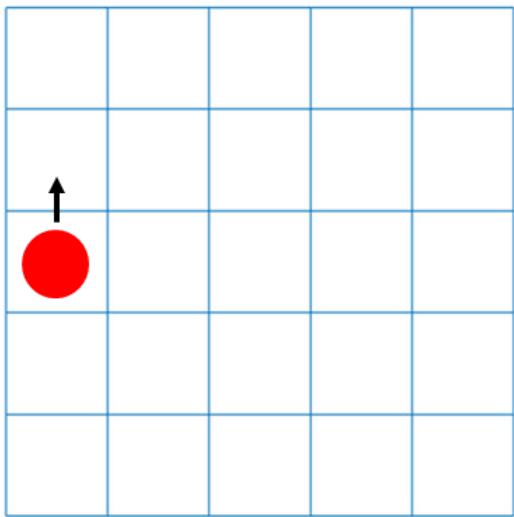
Grid World Model

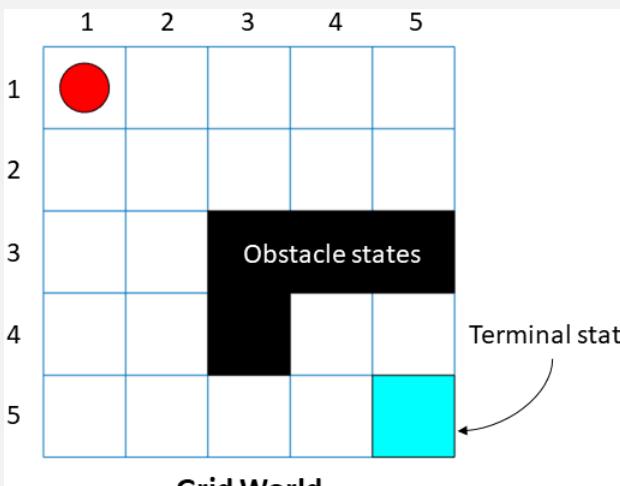
You can create your own grid world model using the `createGridWorld` function. Specify the grid size when creating the `GridWorld` model object.

The `GridWorld` object has the following properties.

Property	Read-Only	Description
GridSize	Yes	Dimensions of the grid world, displayed as an m -by- n array. Here, m represents the number of grid rows and n is the number of grid columns.

Property	Read-Only	Description						
CurrentState	No	<p>Name of the current state of the agent, specified as a string. You can use this property to set the initial state of the agent. The agent always starts from cell [1,1] by default.</p> <p>The agent starts from the <code>CurrentState</code> once you use the reset function in the <code>rlMDPEnv</code> environment object.</p>						
States	Yes	<p>A string vector containing the state names of the grid world. For instance, for a 2-by-2 grid world model <code>GW</code>, specify the following:</p> <pre>GW.States = ["[1,1]"; "[2,1]"; "[1,2]"; "[2,2]"];</pre>						
Actions	Yes	<p>A string vector containing the list of possible actions that the agent can use. You can set the actions when you create the grid world model by using the <code>moves</code> argument:</p> <pre>GW = createGridWorld(m,n,moves)</pre> <p>Specify <code>moves</code> as either 'Standard' or 'Kings'.</p> <table border="1"> <thead> <tr> <th>moves</th> <th>Gw.Actions</th> </tr> </thead> <tbody> <tr> <td>'Standard'</td> <td>['N'; 'S'; 'E'; 'W']</td> </tr> <tr> <td>'Kings'</td> <td>['N'; 'S'; 'E'; 'W'; 'NE'; 'NW'; 'SE'; 'SW']</td> </tr> </tbody> </table>	moves	Gw.Actions	'Standard'	['N'; 'S'; 'E'; 'W']	'Kings'	['N'; 'S'; 'E'; 'W'; 'NE'; 'NW'; 'SE'; 'SW']
moves	Gw.Actions							
'Standard'	['N'; 'S'; 'E'; 'W']							
'Kings'	['N'; 'S'; 'E'; 'W'; 'NE'; 'NW'; 'SE'; 'SW']							

Property	Read-Only	Description																																			
T	No	<p>State transition matrix, specified as a 3-D array. T is a probability matrix that indicates the likelihood of the agent moving from the current state s to any possible next state s' by performing action a.</p>  <p>T can be denoted as</p> $T(s, s', a) = \text{probability}(s' s, a).$ <p>For instance, consider a 5-by-5 deterministic grid world object GW with the agent in cell [3, 1]. View the state transition matrix for the north direction.</p> <pre>northStateTransition = GW.T(:,:,1)</pre> <p>northStateTransition = 25x25</p> <table border="1"> <tr> <td>"[1,1]"</td> <td>"[2,1]"</td> <td>"[3,1]"</td> <td></td> <td></td> </tr> <tr> <td>"[1,1]"</td> <td>1</td> <td>0</td> <td>0</td> <td></td> </tr> <tr> <td>"[2,1]"</td> <td>1</td> <td>0</td> <td>0</td> <td></td> </tr> <tr> <td>"[3,1]"</td> <td>0</td> <td>1</td> <td>0</td> <td></td> </tr> <tr> <td>"[4,1]"</td> <td>0</td> <td>0</td> <td>1</td> <td></td> </tr> <tr> <td>"[5,1]"</td> <td>0</td> <td>0</td> <td>0</td> <td></td> </tr> <tr> <td>"[1,2]"</td> <td>0</td> <td>0</td> <td>0</td> <td></td> </tr> </table>  <p>From the above figure, the value of <code>northStateTransition(3, 2)</code> is 1 since the agent moves from cell [3, 1] to cell [2, 1] with action 'N'. A probability of 1 indicates that from a given state, if the agent goes north, it has a 100% chance of moving one cell north on the grid. For an example showing how to set up the state transition matrix, see "Train Reinforcement Learning Agent in Basic Grid World" on page 1-8.</p>	"[1,1]"	"[2,1]"	"[3,1]"			"[1,1]"	1	0	0		"[2,1]"	1	0	0		"[3,1]"	0	1	0		"[4,1]"	0	0	1		"[5,1]"	0	0	0		"[1,2]"	0	0	0	
"[1,1]"	"[2,1]"	"[3,1]"																																			
"[1,1]"	1	0	0																																		
"[2,1]"	1	0	0																																		
"[3,1]"	0	1	0																																		
"[4,1]"	0	0	1																																		
"[5,1]"	0	0	0																																		
"[1,2]"	0	0	0																																		

Property	Read-Only	Description
R	No	<p>Reward transition matrix, specified as a 3-D array. R determines how much reward the agent receives after performing an action in the environment. R has the same shape and size as the state transition matrix T.</p> <p>The reward transition matrix R can be denoted as</p> $r = R(s, s', a).$ <p>Set up R such that there is a reward to the agent after every action. For instance, you can set up a positive reward if the agent transitions over obstacle states and when it reaches the terminal state. You can also set up a default reward of -11 for all actions the agent takes, independent of the current state and next state. For an example that show how to set up the reward transition matrix, see “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8.</p>
ObstacleStates	No	<p>ObstacleStates are states that cannot be reached in the grid world, specified as a string vector. Consider the following 5-by-5 grid world model GW.</p>  <p>The black cells are obstacle states, and you can specify them using the following syntax:</p> <pre>GW.ObstacleStates = ["[3,3]"; "[3,4]"; "[3,5]"; "[4,3]"];</pre> <p>For a workflow example, see “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8.</p>

Property	Read-Only	Description
TerminalStates	No	<p>TerminalStates are the final states in the grid world, specified as a string vector. Consider the previous 5-by-5 grid world model GW. The blue cell is the terminal state and you can specify it by:</p> <pre>GW.TerminalStates = "[5,5]";</pre> <p>For a workflow example, see “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8.</p>

Grid World Environment

You can create a Markov decision process (MDP) environment using `rlMDPEnv` from the grid world model from the previous step. MDP is a discrete-time stochastic control process. It provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of the decision maker. The agent uses the grid world environment object `rlMDPEnv` to interact with the grid world model object `GridWorld`.

For more information, see `rlMDPEnv` and “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8.

See Also

`createGridWorld` | `rlMDPEnv` | `rlPredefinedEnv`

More About

- “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8

Create MATLAB Environment Using Custom Functions

This example shows how to create a cart-pole environment by supplying custom dynamic functions in MATLAB®.

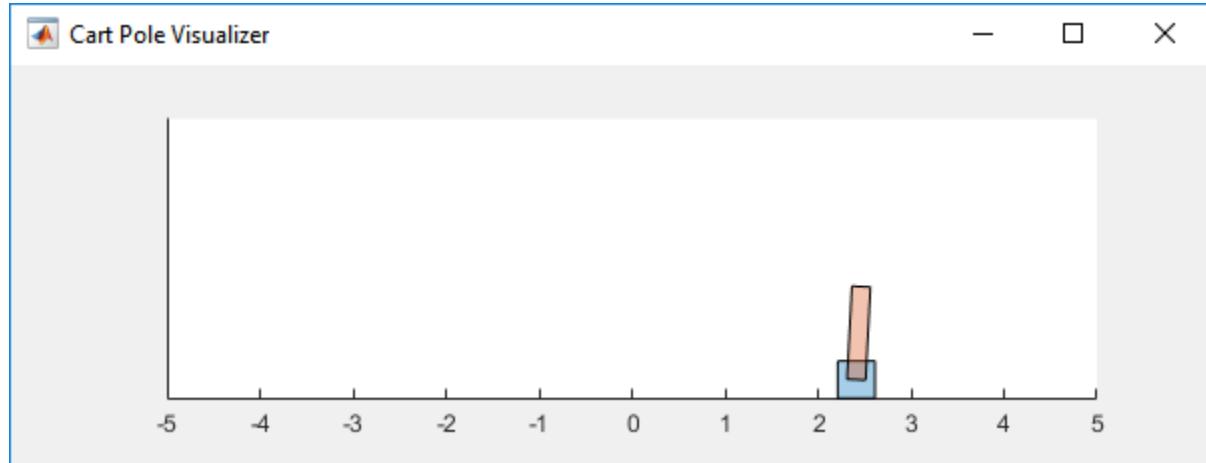
Using the `rlFunctionEnv` function, you can create a MATLAB reinforcement learning environment from an observation specification, an action specification, and user-defined `step` and `reset` functions. You can then train a reinforcement learning agent in this environment. The necessary `step` and `reset` functions are already defined for this example.

Creating an environment using custom functions is useful for environments with less complex dynamics, environments with no special visualization requirements, or environments with interfaces to third-party libraries. For more complex environments, you can create an environment object using a template class. For more information, see “Create Custom MATLAB Environment from Template” on page 2-40.

For more information on creating reinforcement learning environments, see “Create MATLAB Environments for Reinforcement Learning” on page 2-2 and “Create Simulink Environments for Reinforcement Learning” on page 2-5.

Cart-Pole MATLAB Environment

The cart-pole environment is a pole attached to an unactuated joint on a cart, which moves along a frictionless track. The training goal is to make the pendulum stand upright without falling over.



For this environment:

- The upward balanced pendulum position is 0 radians, and the downward hanging position is π radians.
- The pendulum starts upright with an initial angle that is between -0.5 and 0.05 .
- The force action signal from the agent to the environment is from -10 to 10 N.
- The observations from the environment are the cart position, cart velocity, pendulum angle, and pendulum angle derivative.
- The episode terminates if the pole is more than 12 degrees from vertical, or if the cart moves more than 2.4 m from the original position.

- A reward of +1 is provided for every time step that the pole remains upright. A penalty of -10 is applied when the pendulum falls.

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Observation and Action Specifications

The observations from the environment are the cart position, cart velocity, pendulum angle, and pendulum angle derivative.

```
ObservationInfo = rlNumericSpec([4 1]);
ObservationInfo.Name = 'CartPole States';
ObservationInfo.Description = 'x, dx, theta, dtheta';
```

The environment has a discrete action space where the agent can apply one of two possible force values to the cart: -10 or 10 N.

```
ActionInfo = rlFiniteSetSpec([-10 10]);
ActionInfo.Name = 'CartPole Action';
```

For more information on specifying environment actions and observations, see `rlNumericSpec` and `rlFiniteSetSpec`.

Create Environment Using Function Names

To define a custom environment, first specify the custom `step` and `reset` functions. These functions must be in your current working folder or on the MATLAB path.

The custom `reset` function sets the default state of the environment. This function must have the following signature.

```
[InitialObservation,LoggedSignals] = myResetFunction()
```

To pass information from one step to the next, such as the environment state, use `LoggedSignals`. For this example, `LoggedSignals` contains the states of the cart-pole environment, the position and velocity of the cart, the pendulum angle, and the pendulum angle derivative. The `reset` function sets the cart angle to a random value each time the environment is reset.

For this example, use the custom reset function defined in `myResetFunction.m`.

```
type myResetFunction.m

function [InitialObservation, LoggedSignal] = myResetFunction()
% Reset function to place custom cart-pole environment into a random
% initial state.

% Theta (randomize)
T0 = 2 * 0.05 * rand() - 0.05;
% Thetadot
Td0 = 0;
% X
X0 = 0;
% Xdot
Xd0 = 0;

% Return initial environment state variables as logged signals.
```

```

LoggedSignal.State = [X0;Xd0;T0;Td0];
InitialObservation = LoggedSignal.State;
end

```

The custom `step` function specifies how the environment advances to the next state based on a given action. This function must have the following signature.

```
[Observation,Reward,IsDone,LoggedSignals] = myStepFunction(Action,LoggedSignals)
```

To get the new state, the environment applies the dynamic equation to the current state stored in `LoggedSignals`, which is similar to giving an initial condition to a differential equation. The new state is stored in `LoggedSignals` and returned as an output.

For this example, use the custom step function defined in `myStepFunction.m`. For implementation simplicity, this function redefines physical constants, such as the cart mass, every time `step` is executed.

```
type myStepFunction.m
```

```

function [NextObs,Reward,IsDone,LoggedSignals] = myStepFunction(Action,LoggedSignals)
% Custom step function to construct cart-pole environment for the function
% name case.
%
% This function applies the given action to the environment and evaluates
% the system dynamics for one simulation step.

% Define the environment constants.

% Acceleration due to gravity in m/s^2
Gravity = 9.8;
% Mass of the cart
CartMass = 1.0;
% Mass of the pole
PoleMass = 0.1;
% Half the length of the pole
HalfPoleLength = 0.5;
% Max force the input can apply
MaxForce = 10;
% Sample time
Ts = 0.02;
% Pole angle at which to fail the episode
AngleThreshold = 12 * pi/180;
% Cart distance at which to fail the episode
DisplacementThreshold = 2.4;
% Reward each time step the cart-pole is balanced
RewardForNotFalling = 1;
% Penalty when the cart-pole fails to balance
PenaltyForFalling = -10;

% Check if the given action is valid.
if ~ismember(Action,[-MaxForce MaxForce])
    error('Action must be %g for going left and %g for going right.',..., ...
        -MaxForce,MaxForce);
end
Force = Action;

% Unpack the state vector from the logged signals.

```

```
State = LoggedSignals.State;
XDot = State(2);
Theta = State(3);
ThetaDot = State(4);

% Cache to avoid recompilation.
CosTheta = cos(Theta);
SinTheta = sin(Theta);
SystemMass = CartMass + PoleMass;
temp = (Force + PoleMass*HalfPoleLength*ThetaDot*ThetaDot*SinTheta)/SystemMass;

% Apply motion equations.
ThetaDotDot = (Gravity*SinTheta - CosTheta*temp) / ...
    (HalfPoleLength*(4.0/3.0 - PoleMass*CosTheta*CosTheta/SystemMass));
XDotDot = temp - PoleMass*HalfPoleLength*ThetaDotDot*CosTheta/SystemMass;

% Perform Euler integration.
LoggedSignals.State = State + Ts.*[XDot;XDotDot;ThetaDot;ThetaDotDot];

% Transform state to observation.
NextObs = LoggedSignals.State;

% Check terminal condition.
X = NextObs(1);
Theta = NextObs(3);
IsDone = abs(X) > DisplacementThreshold || abs(Theta) > AngleThreshold;

% Get reward.
if ~IsDone
    Reward = RewardForNotFalling;
else
    Reward = PenaltyForFalling;
end
end
```

Construct the custom environment using the defined observation specification, action specification, and function names.

```
env = rlFunctionEnv(ObservationInfo, ActionInfo, 'myStepFunction', 'myResetFunction');
```

To verify the operation of your environment, `rlFunctionEnv` automatically calls `validateEnvironment` after creating the environment.

Create Environment Using Function Handles

You can also define custom functions that have additional input arguments beyond the minimum required set. For example, to pass the additional arguments `arg1` and `arg2` to both the step and rest function, use the following code.

```
[InitialObservation,LoggedSignals] = myResetFunction(arg1,arg2)
[Observation,Reward,IsDone,LoggedSignals] = myStepFunction(Action,LoggedSignals,arg1,arg2)
```

To use these functions with `rlFunctionEnv`, you must use anonymous function handles.

```
ResetHandle = @myResetFunction(arg1,arg2);
StepHandle = @(Action,LoggedSignals) myStepFunction(Action,LoggedSignals,arg1,arg2);
```

For more information, see “Anonymous Functions” (MATLAB).

Using additional input arguments can create a more efficient environment implementation. For example, `myStepFunction2.m` contains a custom `step` function that takes the environment constants as an input argument (`envConstants`). By doing so, this function avoids redefining the environment constants at each step.

```

type myStepFunction2.m

function [NextObs,Reward,IsDone,LoggedSignals] = myStepFunction2(Action,LoggedSignals,EnvConstan
% Custom step function to construct cart-pole environment for the function
% handle case.
%
% This function applies the given action to the environment and evaluates
% the system dynamics for one simulation step.

% Check if the given action is valid.
if ~ismember(Action,[-EnvConstants.MaxForce EnvConstants.MaxForce])
    error('Action must be %g for going left and %g for going right.',..., ...
        -EnvConstants.MaxForce,EnvConstants.MaxForce);
end
Force = Action;

% Unpack the state vector from the logged signals.
State = LoggedSignals.State;
XDot = State(2);
Theta = State(3);
ThetaDot = State(4);

% Cache to avoid recomputation.
CosTheta = cos(Theta);
SinTheta = sin(Theta);
SystemMass = EnvConstants.MassCart + EnvConstants.MassPole;
temp = (Force + EnvConstants.MassPole*EnvConstants.Length*ThetaDot*ThetaDot*SinTheta)/SystemMass

% Apply motion equations.
ThetaDotDot = (EnvConstants.Gravity*SinTheta - CosTheta*temp)...
    / (EnvConstants.Length*(4.0/3.0 - EnvConstants.MassPole*CosTheta*CosTheta/SystemMass));
XDotDot = temp - EnvConstants.MassPole*EnvConstants.Length*ThetaDotDot*CosTheta/SystemMass;

% Perform Euler integration.
LoggedSignals.State = State + EnvConstants.Ts.*[XDot;XDotDot;ThetaDot;ThetaDotDot];

% Transform state to observation.
NextObs = LoggedSignals.State;

% Check terminal condition.
X = NextObs(1);
Theta = NextObs(3);
IsDone = abs(X) > EnvConstants.XThreshold || abs(Theta) > EnvConstants.ThetaThresholdRadians;

% Get reward.
if ~IsDone
    Reward = EnvConstants.RewardForNotFalling;
else
    Reward = EnvConstants.PenaltyForFalling;
end

end

```

Create the structure that contains the environment constants.

```
% Acceleration due to gravity in m/s^2
envConstants.Gravity = 9.8;
% Mass of the cart
envConstants.MassCart = 1.0;
% Mass of the pole
envConstants.MassPole = 0.1;
% Half the length of the pole
envConstants.Length = 0.5;
% Max force the input can apply
envConstants.MaxForce = 10;
% Sample time
envConstants.Ts = 0.02;
% Angle at which to fail the episode
envConstants.ThetaThresholdRadians = 12 * pi/180;
% Distance at which to fail the episode
envConstants.XThreshold = 2.4;
% Reward each time step the cart-pole is balanced
envConstants.RewardForNotFalling = 1;
% Penalty when the cart-pole fails to balance
envConstants.PenaltyForFalling = -5;
```

Create an anonymous function handle to the custom `step` function, passing `envConstants` as an additional input argument. Because `envConstants` is available at the time that `StepHandle` is created, the function handle includes those values. The values persist within the function handle even if you clear the variables.

```
StepHandle = @(Action,LoggedSignals) myStepFunction2(Action,LoggedSignals,envConstants);
```

Use the same `reset` function, specifying it as a function handle rather than by using its name.

```
ResetHandle = @myResetFunction;
```

Create the environment using the custom function handles.

```
env2 = rlFunctionEnv(ObservationInfo,ActionInfo,StepHandle,ResetHandle);
```

Validate Custom Functions

Before you train an agent in your environment, the best practice is to validate the behavior of your custom functions. To do so, you can initialize your environment using the `reset` function and run one simulation step using the `step` function. For reproducibility, set the random generator seed before validation.

Validate the environment created using function names.

```
rng(0);
InitialObs = reset(env)

InitialObs = 4x1

    0
    0
    0.0315
    0
```

```
[NextObs,Reward,IsDone,LoggedSignals] = step(env,10);
NextObs
NextObs = 4x1
          0
    0.1947
    0.0315
   -0.2826
```

Validate the environment created using function handles.

```
rng(0);
InitialObs2 = reset(env2)

InitialObs2 = 4x1
          0
          0
    0.0315
          0

[NextObs2,Reward2,IsDone2,LoggedSignals2] = step(env2,10);
NextObs
NextObs = 4x1
          0
    0.1947
    0.0315
   -0.2826
```

Both environments initialize and simulate successfully, producing the same state values in `NextObs`.

See Also

`rlFunctionEnv`

More About

- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Create Custom MATLAB Environment from Template” on page 2-40

Create Custom MATLAB Environment from Template

You can define a custom reinforcement learning environment by creating and modifying a template environment class. You can use a custom template environment to:

- Implement more complex environment dynamics.
- Add custom visualizations to your environment.
- Create an interface to third-party libraries defined in languages such as C++, Java®, or Python®.
For more information, see “External Language Interfaces” (MATLAB).

For more information about creating MATLAB classes, see “User-Defined Classes” (MATLAB).

You can create less complex custom reinforcement learning environments using custom functions, as described in “Create MATLAB Environment Using Custom Functions” on page 2-33.

Create Template Class

To define your custom environment, first create the template class file, specifying the name of the class. For this example, name the class `MyEnvironment`.

```
rlCreateEnvTemplate("MyEnvironment")
```

The software creates and opens the template class file. The template class is a subclass of the `rl.env.MATLABEnvironment` abstract class, as shown in the class definition at the start of the template file. This abstract class is the same one used by the other MATLAB reinforcement learning environment objects.

```
classdef MyEnvironment < rl.env.MATLABEnvironment
```

By default, the template class implements a simple cart-pole balancing model similar to the cart-pole predefined environments described in “Load Predefined Control System Environments” on page 2-15.

To define your environment dynamics modify the template class, specify the following:

- Environment properties
- Required environment methods
- Optional environment methods

Environment Properties

In the `properties` section of the template, specify any parameters necessary for creating and simulating the environment. These parameters can include:

- Physical constants — The sample environment defines the acceleration due to gravity (`Gravity`).
- Environment geometry — The sample environment defines the cart and pole masses (`CartMass` and `PoleMass`) and the half-length of the pole (`HalfPoleLength`).
- Environment constraints — The sample environment defines the pole angle and cart distance thresholds (`AngleThreshold` and `DisplacementThreshold`). The environment uses these values to detect when a training episode is finished.
- Variables required for evaluating the environment — The sample environment defines the state vector (`State`) and a flag for indicating when an episode is finished (`IsDone`).

- Constants for defining the actions or observation spaces — The sample environment defines the maximum force for the action space (`MaxForce`).
- Constants for calculating the reward signal — The sample environment defines the constants `RewardForNotFalling` and `PenaltyForFalling`.

```

properties
    % Specify and initialize the necessary properties of the environment
    % Acceleration due to gravity in m/s^2
    Gravity = 9.8

    % Mass of the cart
    CartMass = 1.0

    % Mass of the pole
    PoleMass = 0.1

    % Half the length of the pole
    HalfPoleLength = 0.5

    % Max force the input can apply
    MaxForce = 10

    % Sample time
    Ts = 0.02

    % Angle at which to fail the episode (radians)
    AngleThreshold = 12 * pi/180

    % Distance at which to fail the episode
    DisplacementThreshold = 2.4

    % Reward each time step the cart-pole is balanced
    RewardForNotFalling = 1

    % Penalty when the cart-pole fails to balance
    PenaltyForFalling = -10
end

properties
    % Initialize system state [x,dx,theta,dtheta]'
    State = zeros(4,1)
end

properties(Access = protected)
    % Initialize internal flag to indicate episode termination
    IsDone = false
end

```

Required Functions

A reinforcement learning environment requires the following functions to be defined. The `getObservationInfo`, `getActionInfo`, `sim`, and `validateEnvironment` functions are already defined in the base abstract class. To create your environment, you must define the constructor, `reset`, and `step` functions.

Function	Description
<code>getObservationInfo</code>	Return information about the environment observations
<code>getActionInfo</code>	Return information about the environment actions
<code>sim</code>	Simulate the environment with an agent
<code>validateEnvironment</code>	Validate the environment by calling the <code>reset</code> function and simulating the environment for one time step using <code>step</code>
<code>reset</code>	Initialize the environment state and clean up any visualization

Function	Description
step	Apply an action, simulate the environment for one step, and output the observations and rewards; also, set a flag indicating whether the episode is complete
Constructor function	A function with the same name as the class that creates an instance of the class

Sample Constructor Function

The sample cart-pole constructor function creates the environment by:

- Defining the action and observation specifications. For more information about creating these specifications, see `rlNumericSpec` and `rlFiniteSetSpec`.
- Calling the constructor of the base abstract class.

```
function this = MyEnvironment()
    % Initialize observation settings
    ObservationInfo = rlNumericSpec([4 1]);
    ObservationInfo.Name = 'CartPole States';
    ObservationInfo.Description = 'x, dx, theta, dtheta';

    % Initialize action settings
    ActionInfo = rlFiniteSetSpec([-1 1]);
    ActionInfo.Name = 'CartPole Action';

    % The following line implements built-in functions of the RL environment
    this = this@rl.env.MATLABEnvironment(ObservationInfo,ActionInfo);

    % Initialize property values and precompute necessary values
    updateActionInfo(this);
end
```

This sample constructor function does not include any input arguments. However, you can add input arguments for your custom constructor.

Sample reset Function

The sample cart-pole reset function sets the initial condition of the model and returns the initial values of the observations. It also generates a notification that the environment has been updated by calling the `envUpdatedCallback` function, which is useful for updating the environment visualization.

```
% Reset environment to initial state and return initial observation
function InitialObservation = reset(this)
    % Theta (+-.05 rad)
    T0 = 2 * 0.05 * rand - 0.05;
    % Thetadot
    Td0 = 0;
    % X
    X0 = 0;
    % Xdot
    Xd0 = 0;

    InitialObservation = [X0;Xd0;T0;Td0];
    this.State = InitialObservation;

    % (Optional) Use notifyEnvUpdated to signal that the
    % environment is updated (for example, to update the visualization)
    notifyEnvUpdated(this);
end
```

Sample step Function

The sample cart-pole `step` function:

- Processes the input action.
- Evaluates the environment dynamic equations for one time step.
- Computes and returns the updated observations.
- Computes and returns the reward signal.
- Checks if the episode is complete and returns the `IsDone` signal as appropriate.
- Generates a notification that the environment has been updated.

```
function [Observation,Reward,IsDone,LoggedSignals] = step(this,Action)
    LoggedSignals = [];

    % Get action
    Force = getForce(this,Action);

    % Unpack state vector
    XDot = this.State(2);
    Theta = this.State(3);
    ThetaDot = this.State(4);

    % Cache to avoid recomputation
    CosTheta = cos(Theta);
    SinTheta = sin(Theta);
    SystemMass = this.CartMass + this.PoleMass;
    temp = (Force + this.PoleMass*this.HalfPoleLength*ThetaDot^2*SinTheta) ...
        /SystemMass;

    % Apply motion equations
    ThetaDotDot = (this.Gravity*SinTheta - CosTheta*temp)...
        / (this.HalfPoleLength*(4.0/3.0 - this.PoleMass*CosTheta*CosTheta/SystemMass));
    XDotDot = temp - this.PoleMass*this.HalfPoleLength*ThetaDotDot*CosTheta/SystemMass;

    % Euler integration
    Observation = this.State + this.Ts.*[XDot;XDotDot;ThetaDot;ThetaDotDot];

    % Update system states
    this.State = Observation;

    % Check terminal condition
    X = Observation(1);
    Theta = Observation(3);
    IsDone = abs(X) > this.DisplacementThreshold || abs(Theta) > this.AngleThreshold;
    this.IsDone = IsDone;

    % Get reward
    Reward = getReward(this);

    % (Optional) Use notifyEnvUpdated to signal that the
    % environment has been updated (for example, to update the visualization)
    notifyEnvUpdated(this);
end
```

Optional Functions

You can define any other functions in your template class as required. For example, you can create helper functions that are called by either `step` or `reset`. The cart-pole template model implements a `getReward` function for computing the reward at each time step.

```
function Reward = getReward(this)
    if ~this.IsDone
        Reward = this.RewardForNotFalling;
```

```

    else
        Reward = this.PenaltyForFalling;
    end
end

```

Environment Visualization

You can add a visualization to your custom environment by implementing the `plot` function. In the `plot` function:

- Create a figure or an instance of a visualizer class of your own implementation. For this example, you create a figure and store a handle to the figure within the environment object.
- Call the `envUpdatedCallback` function.

```

function plot(this)
    % Initiate the visualization
    this.Figure = figure('Visible','on','HandleVisibility','off');
    ha = gca(this.Figure);
    ha.XLimMode = 'manual';
    ha.YLimMode = 'manual';
    ha.XLim = [-3 3];
    ha.YLim = [-1 2];
    hold(ha,'on');
    % Update the visualization
    envUpdatedCallback(this)
end

```

For this example, store the handle to the figure as a protected property of the environment object.

```

properties(Access = protected)
    % Initialize internal flag to indicate episode termination
    IsDone = false

    % Handle to figure
    Figure
end

```

In the `envUpdatedCallback`, plot the visualization to the figure or use your custom visualizer object. For example, check if the figure handle has been set. If it has, then plot the visualization.

```

function envUpdatedCallback(this)
    if isempty(this.Figure) && isvalid(this.Figure)
        % Set visualization figure as the current figure
        ha = gca(this.Figure);

        % Extract the cart position and pole angle
        x = this.State(1);
        theta = this.State(3);

        cartplot = findobj(ha,'Tag','cartplot');
        poleplot = findobj(ha,'Tag','poleplot');
        if isempty(cartplot) || ~isvalid(cartplot) ...
            || isempty(poleplot) || ~isvalid(poleplot)
            % Initialize the cart plot
            cartpoly = polyshape([-0.25 -0.25 0.25 0.25],[-0.125 0.125 0.125 -0.125]);
            cartpoly = translate(cartpoly,[x 0]);
            cartplot = plot(ha,cartpoly,'FaceColor',[0.8500 0.3250 0.0980]);
            cartplot.Tag = 'cartplot';

            % Initialize the pole plot
            L = this.HalfPoleLength*2;
    end

```

```

polepoly = polyshape([-0.1 -0.1 0.1 0.1],[0 L L 0]);
polepoly = translate(polepoly,[x,0]);
polepoly = rotate(polepoly,rad2deg(theta),[x,0]);
poleplot = plot(ha,polepoly,'FaceColor',[0 0.4470 0.7410]);
poleplot.Tag = 'poleplot';
else
    cartpoly = cartplot.Shape;
    polepoly = poleplot.Shape;
end

% Compute the new cart and pole position
[cartposx,~] = centroid(cartpoly);
[poleposx,poleposy] = centroid(polepoly);
dx = x - cartposx;
dtheta = theta - atan2(cartposx-poleposx,poleposy-0.25/2);
cartpoly = translate(cartpoly,[dx,0]);
polepoly = translate(polepoly,[dx,0]);
polepoly = rotate(polepoly,rad2deg(dtheta),[x,0.25/2]);

% Update the cart and pole positions on the plot
cartplot.Shape = cartpoly;
poleplot.Shape = polepoly;

% Refresh rendering in the figure window
drawnow();
end
end

```

The environment calls the `envUpdatedCallback` function, and therefore updates the visualization, whenever the environment is updated.

Create Custom Environment

After you define your custom environment class, create an instance of it in the MATLAB workspace. At the command line, type the following.

```
env = MyEnvironment;
```

If your constructor has input arguments, specify them after the class name. For example, `MyEnvironment(arg1,arg2)`.

After you create your environment, the best practice is to validate the environment dynamics. To do so, use the `validateEnvironment` function, which prints an error to the command window if your environment implementation has any issues.

```
validateEnvironment(env)
```

After validating the environment object, you can use it to train a reinforcement learning agent. For more information on training agents, see “Train Reinforcement Learning Agents” on page 5-2.

See Also

`rlCreateEnvTemplate` | `train`

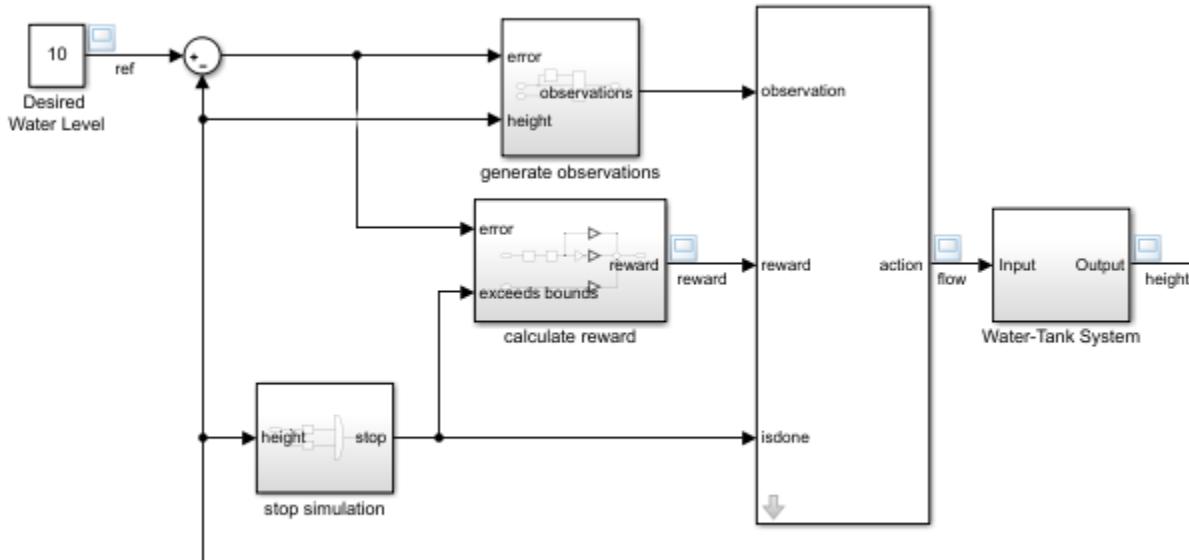
More About

- “Create MATLAB Environments for Reinforcement Learning” on page 2-2
- “Create MATLAB Environment Using Custom Functions” on page 2-33
- “Define Reward Signals” on page 2-7

Water Tank Reinforcement Learning Environment Model

This example shows how to create a water tank reinforcement learning Simulink® environment that contains an RL Agent block in the place of a controller for the water level in a tank. To simulate this environment, you must create an agent and specify that agent in the RL Agent block. For an example that trains an agent using this environment, see “Create Simulink Environment and Train Agent” on page 1-19.

```
mdl = 'rlwatertank';
open_system(mdl)
```



This model already contains an RL Agent block, which connects to the following signals:

- Scalar action output signal
- Vector of observation input signals
- Scalar reward input signal
- Logical input signal for stopping the simulation

Actions and Observations

A reinforcement learning environment receives action signals from the agent and generates observation signals in response to these actions. To create and train an agent, you must create action and observation specification objects.

The action signal for this environment is the flow rate control signal that is sent to the plant. To create a specification object for this continuous action signal, use the `rlNumericSpec` function.

```
actionInfo = rlNumericSpec([1 1]);
actionInfo.Name = 'flow';
```

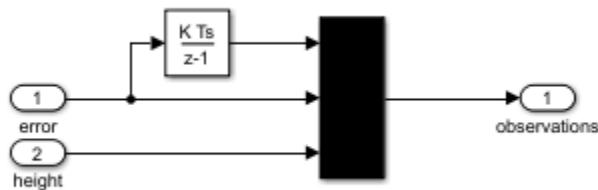
If the action signal takes one of a discrete set of possible values, create the specification using the `rlFiniteSetSpec` function.

For this environment, there are three observation signals sent to the agent, specified as a vector signal. The observation vector is $[f e dt e h]^T$, where:

- h is the height of the water in the tank.
- $e = r - h$, where r is the reference value for the water height.

Compute the observation signals in the generate observations subsystem.

```
open_system([mdl '/generate_observations'])
```



Create a three-element vector of observation specifications. Specify a lower bound of 0 for the water height, leaving the other observation signals unbounded.

```
observationInfo = rlNumericSpec([3 1],...
    'LowerLimit', [-inf -inf 0],...
    'UpperLimit', [ inf inf inf]);
observationInfo.Name = 'observations';
observationInfo.Description = 'integrated error, error, and measured height';
```

If the actions or observations are represented by bus signals, create specifications using the `bus2RLSpec` function.

Reward Signal

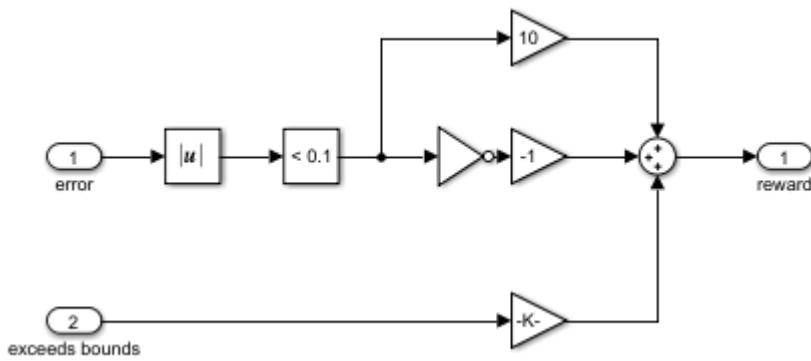
Construct a scalar reward signal. For this example, specify the following reward.

$$\text{reward} = 10(|e| < 0.1) - 1(|e| \geq 0.1) - 100(h \leq 0 \mid h \geq 20)$$

The reward is positive when the error is below 0.1 and negative otherwise. Also, there is a large reward penalty when the water height is outside the 0 to 20 range.

Construct this reward in the calculate reward subsystem.

```
open_system([mdl '/calculate_reward'])
```

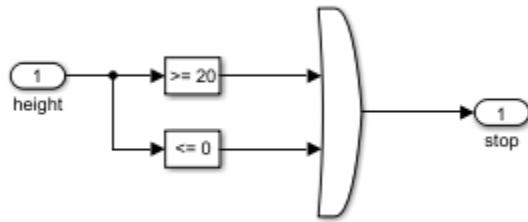


Stop Signal

To terminate training episodes and simulations, specify a logical signal to the `isdone` input port of the block. For this example, terminate the episode if $h \leq 0$ or $h \geq 20$.

Compute this signal in the stop simulation subsystem.

```
open_system([mdl '/stop simulation'])
```



Create Environment Object

Create an environment object for the Simulink model.

```
env = rlSimulinkEnv(mdl,[mdl '/RL Agent'],observationInfo,actionInfo);
```

Reset Function

You can also create a custom reset function that randomizes parameters, variables, or states of the model. In this example, the reset function randomizes the reference signal and the initial water height and sets the corresponding block parameters.

```
env.ResetFcn = @(in)localResetFcn(in);
```

Local Function

```
function in = localResetFcn(in)

% Randomize reference signal
blk = sprintf('rlwatertank/Desired \nWater Level');
h = 3*randn + 10;
while h <= 0 || h >= 20
```

```
    h = 3*randn + 10;
end
in = setBlockParameter(in,blk,'Value',num2str(h));

% Randomize initial height
h = 3*randn + 10;
while h <= 0 || h >= 20
    h = 3*randn + 10;
end
blk = 'rlwatertank/Water-Tank System/H';
in = setBlockParameter(in,blk,'InitialCondition',num2str(h));

end
```

See Also

[rlSimulinkEnv](#)

More About

- “Create Simulink Environments for Reinforcement Learning” on page 2-5

Define Policies and Value Functions

- “Create Policy and Value Function Representations” on page 3-2
- “Import Policy and Value Function Representations” on page 3-9

Create Policy and Value Function Representations

A reinforcement learning policy is a mapping that selects an action to take based on observations from the environment. During training, the agent tunes the parameters of its policy representation to maximize the long-term reward.

Depending on the type of reinforcement learning agent you are using, you define actor and critic function approximators, which the agent uses to represent and train its policy. The actor represents the policy that selects the best action to take. The critic represents the value function that estimates the long-term reward for the current policy. Depending on your application and selected agent, you can define policy and value functions using deep neural networks, linear basis functions, or lookup tables.

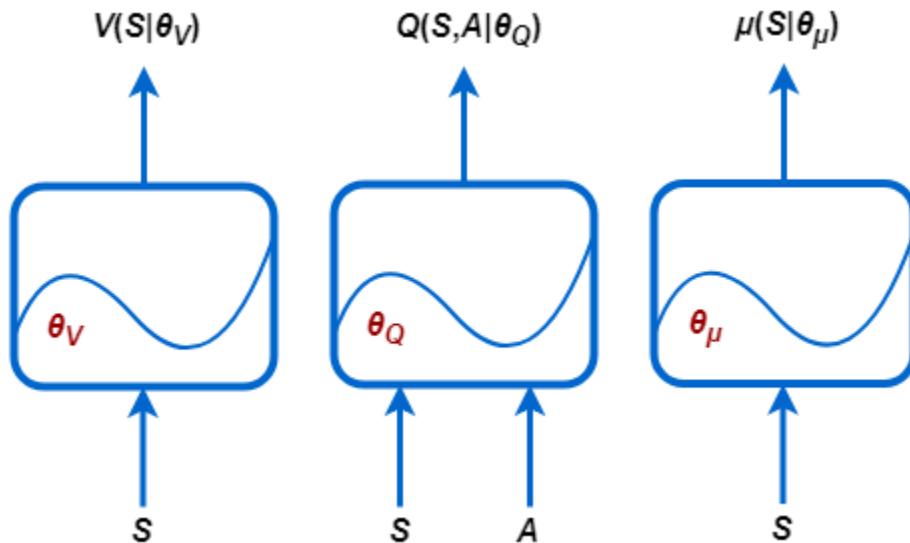
For more information on agents, see “Reinforcement Learning Agents” on page 4-2.

Function Approximation

Depending on the type of agent you are using, Reinforcement Learning Toolbox software supports the following types of function approximators:

- $V(S|\theta_V)$ — Critics that estimate the expected long-term reward based on a given observation S
- $Q(S,A|\theta_Q)$ — Critics that estimate the expected long-term reward based on a given observation S and action A
- $Q(S,A_i|\theta_Q)$ — Critics that estimate the expected long-term reward for all possible discrete actions given observation S
- $\mu(S|\theta_\mu)$ — Actors that select an action based on a given observation S

Each function approximator has a corresponding set of parameters (θ_V , θ_Q , θ_μ), which are computed during the learning process.



For systems with a limited number of discrete observations and discrete actions, you can store value functions in a lookup table. For systems that have many discrete observations and actions and for

observation and action spaces that are continuous, storing the observations and actions is impractical. For such systems, you can represent your actors and critics using deep neural networks or linear basis functions.

Table Representations

You can create two types of table representations:

- Value tables, which store rewards for corresponding observations
- Q-tables, which store rewards for corresponding observation-action pairs

To create a table representation, first create a value table or Q table using the `rlTable` function. Then, create a representation for the table using either an `rlValueRepresentation` or `rlQValueRepresentation` object. To configure the learning rate and optimization used by the representation, use an `rlRepresentationOptions` object.

Deep Neural Network Representations

You can create actor and critic function approximators using deep neural network representations. Doing so uses Deep Learning Toolbox software features.

Network Input and Output Dimensions

The dimensions of your actor and critic networks must match the corresponding action and observation specifications from the training environment object. To obtain the action and observation dimensions for environment `env`, use the `getActionInfo` and `getObservationInfo` functions, respectively. Then access the `Dimensions` property of the specification objects.

```
actInfo = getActionInfo(env);
actDimensions = actInfo.Dimensions;

obsInfo = getObservationInfo(env);
obsDimensions = obsInfo.Dimensions;
```

For critic networks that take only observations as inputs, such as those used in AC or PG agents, the dimensions of the input layers must match the dimensions of the environment observation specifications. The dimensions of the critic output layer must be a scalar value function.

For critic networks that take both observations and actions as inputs, such as those used in DQN or DDPG agents, the dimensions of the input layers must match the dimensions of the corresponding environment observation and action specifications.

For actor networks the dimensions of the input layers must match the dimensions of the environment observation specifications. If the actor has a:

- Discrete action space, then its output size must equal the number of discrete actions.
- Continuous action space, then its output size must be a scalar or vector value, as defined in the action specification.

Build Deep Neural Network

Deep neural networks consist of a series of interconnected layers. The following table lists some common deep learning layers used in reinforcement learning applications. For a full list of available layers, see “List of Deep Learning Layers” (Deep Learning Toolbox).

Layer	Description
imageInputLayer	Input vectors and 2-D images, and normalize the data.
tanhLayer	Apply a hyperbolic tangent activation layer to the layer inputs.
reluLayer	Set any input values that are less than zero to zero.
fullyConnectedLayer	Multiply the input vector by a weight matrix, and add a bias vector.
convolution2dLayer	Apply sliding convolutional filters to the input.
additionLayer	Add the outputs of multiple layers together.
concatenationLayer	Concatenate inputs along a specified dimension.

The `bilstmLayer`, and `batchNormalizationLayer` layers are not supported for reinforcement learning.

You can also create your own custom layers. For more information, see “Define Custom Deep Learning Layers” (Deep Learning Toolbox). Reinforcement Learning Toolbox software provides the following custom layers.

Layer	Description
scalingLayer	Linearly scale and bias an input array. This layer is useful for scaling and shifting the outputs of nonlinear layers, such as <code>tanhLayer</code> and <code>sigmoid</code> .
quadraticLayer	Create vector of quadratic monomials constructed from the elements of the input array. This layer is useful when you need an output that is some quadratic function of its inputs, such as for an LQR controller.
softplusLayer	Implement the softplus activation $Y = \log(1 + e^X)$, which ensures that the output is always positive

The custom layers do not contain tunable parameters; that is, they do not change during training.

For reinforcement learning applications, you construct your deep neural network by connecting a series of layers for each input path (observations or actions) and for each output path (estimated rewards or actions). You then connect these paths together using the `connectLayers` function.

You can also create your deep neural network using the **Deep Network Designer** app. For an example, see “Create Agent Using Deep Network Designer and Train Using Image Observations” on page 5-72.

When you create a deep neural network, you must specify names for the first layer of each input path and the final layer of the output path.

The following code creates and connects the following input and output paths:

- An observation input path, `observationPath`, with the first layer named '`observation`'.
- An action input path, `actionPath`, with the first layer named '`action`'.
- An estimated value function output path, `commonPath`, which takes the outputs of `observationPath` and `actionPath` as inputs. The final layer of this path is named '`output`'.

```

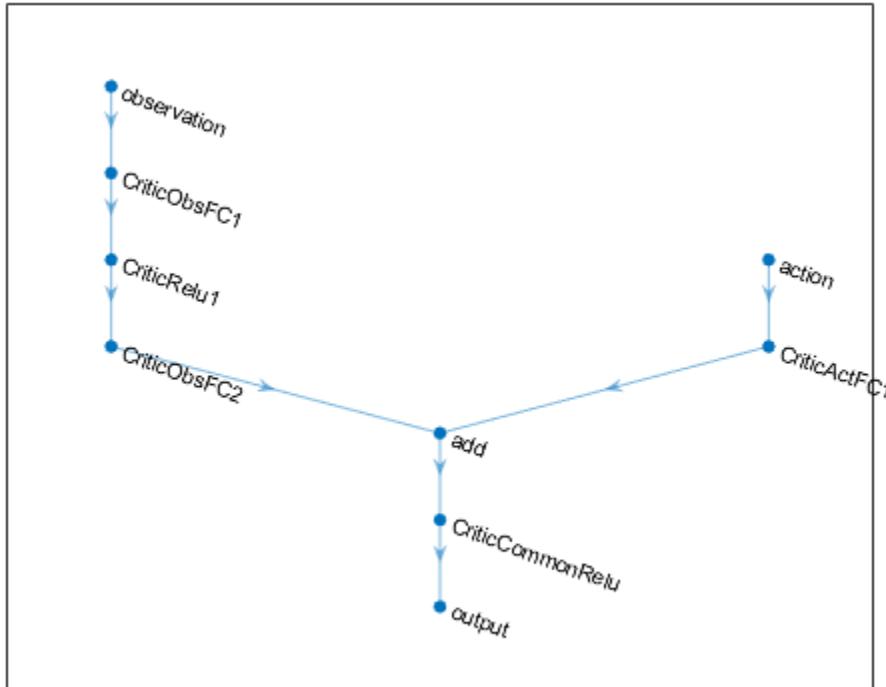
observationPath = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(24, 'Name', 'CriticObsFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(24, 'Name', 'CriticObsFC2'));
actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(24, 'Name', 'CriticActFC1'));
commonPath = [
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'CriticCommonRelu')
    fullyConnectedLayer(1, 'Name', 'output')];
criticNetwork = layerGraph(observationPath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = addLayers(criticNetwork, commonPath);
criticNetwork = connectLayers(criticNetwork, 'CriticObsFC2', 'add/in1');
criticNetwork = connectLayers(criticNetwork, 'CriticActFC1', 'add/in2');

```

For all observation and action input paths, you must specify an `imageInputLayer` as the first layer in the path.

You can view the structure of your deep neural network using the `plot` function.

```
plot(criticNetwork)
```



For PG and AC agents, the final output layers of your deep neural network actor representation are a `fullyConnectedLayer` and a `softmaxLayer` layer. When you specify the layers for your network, you must specify the `fullyConnectedLayer` and you can optionally specify the `softmaxLayer`. If you omit the `softmaxLayer`, the software automatically adds one for you.

Determining the number, type, and size of layers for your deep neural network representation can be difficult and is application dependent. However, the most critical component for any function approximator is whether the function is able to approximate the optimal policy or discounted value

function for your application; that is, whether it has layers that can correctly learn the features of your observation, action, and reward signals.

Consider the following tips when constructing your network.

- For continuous action spaces, bound actions with a `tanhLayer` followed by `ScalingLayer`, if necessary.
- Deep dense networks with `reluLayer` layers can be fairly good at approximating many different functions. Therefore, they are often a good first choice.
- When you approximate strong nonlinearities or systems with algebraic constraints, adding more layers is often better than increasing the number of outputs per layer. Adding more layers promotes exponential exploration, while adding layer outputs promotes polynomial exploration.
- For on-policy agents, such as AC and PG agents, parallel training works better if your networks are large (for example, a network with two hidden layers with 32 nodes each, which has a few hundred parameters). On-policy parallel updates assume each worker updates a different part of the network, such as when they explore different areas of the observation space. If the network is small, the worker updates can correlate with each other and make training unstable.

Create and Configure Representation

To create a critic representation for your deep neural network, use an `rlValueRepresentation` or `rlQValueRepresentation` object. To create an actor representation for your deep neural network, use an `rlDeterministicActorRepresentation` or `rlStochasticActorRepresentation` object. To configure the learning rate and optimization used by the representation, use an `rlRepresentationOptions` object.

For example, create a Q-value representation object for the critic network `criticNetwork`, specifying a learning rate of `0.0001`. When you create the representation, pass the environment action and observation specifications to the `rlQValueRepresentation` object, and specify the names of the network layers to which the actions and observations are connected.

```
opt = rlRepresentationOptions('LearnRate',0.0001);
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo, ...
    'Observation',{'observation'},'Action',{'action'},opt);
```

When you create your deep neural network and configure your representation object, consider using one of the following approaches as a starting point.

- 1 Start with the smallest possible network and a high learning rate (`0.01`). Train this initial network to see if the agent converges quickly to a poor policy or acts in a random manner. If either of these issues occur, rescale the network by adding more layers or more outputs on each layer. Your goal is to find a network structure that is just big enough, does not learn too fast, and shows signs of learning (an improving trajectory of the reward graph) after an initial training period.
- 2 A low initial learning rate can allow you to see if the agent is on the right track, and help you check that your network architecture is satisfactory for the problem. For difficult problems, tuning parameters is much easier once you settle on a good network architecture.

Also, consider the following tips when configuring your deep neural network representation.

- Be patient with DDPG and DQN agents, since they might not learn anything for some time during the early episodes, and they typically show a dip in cumulative reward early in the training process. Eventually, they can show signs of learning after the first few thousand episodes.

- For DDPG and DQN agents, promoting exploration of the agent is critical.
- For agents with both actor and critic networks, set the initial learning rates of both representations to the same value. For some problems, setting the critic learning rate to a higher value than that of the actor can improve learning results.

Recurrent Neural Networks

When creating representations for use with a PPO or DQN agent, you can use recurrent neural networks. These networks are deep neural networks with at least one layer that has hidden state information, such as an `lstmLayer`. For more information and examples, see `rlValueRepresentation`, `rlQValueRepresentation`, `rlDeterministicActorRepresentation`, and `rlStochasticActorRepresentation`.

Linear Basis Function Representations

Linear basis function representations have the form $f = W'B$, where W is a weight array and B is the column vector output of a custom basis function. The learnable parameters of a linear basis function representation are the elements of W .

For critic representations, f is a scalar value and W is a column vector with the same length as B .

For actor representations with a :

- Continuous action space, the dimensions of f match the dimensions of the agent action specification, which is either a scalar or a column vector.
- Discrete action space, f is a column vector with length equal to the number of discrete actions.

For actor representations, the number of columns in W equals the number of elements in f .

To create a linear basis function representation, first create a custom basis function that returns a column vector. The signature of the basis function depends on the type of agent you are creating. For more information, `rlValueRepresentation`, `rlQValueRepresentation`, `rlDeterministicActorRepresentation`, and `rlStochasticActorRepresentation`.

For an example that trains a custom agent that uses a linear basis function representation, see “Train Custom LQR Agent” on page 5-153.

Specify Agent Representations

Once you create your actor and critic representations, you can create a reinforcement learning agent that uses these representations. For example, create a PG agent using a given actor and critic network.

```
agentOpts = rlPGAgentOptions('UseBaseline',true);
agent = rlPGAgent(actor,baseline,agentOpts);
```

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

You can obtain the actor and critic representations from an existing agent using `getActor` and `getCritic`, respectively.

You can also set the actor and critic of an existing agent using `setActor` and `setCritic`, respectively. When you specify a representation using these functions, the input and output layers of

the specified representation must match the observation and action specifications of the original agent.

See Also

More About

- “Reinforcement Learning Agents” on page 4-2
- “Import Policy and Value Function Representations” on page 3-9

Import Policy and Value Function Representations

To create function approximators for reinforcement learning, you can import pretrained deep neural networks or deep neural network layer architectures using the Deep Learning Toolbox network import functionality. You can import:

- Open Neural Network Exchange (ONNX) models, which require the Deep Learning Toolbox Converter for ONNX Model Format support package software. For more information, see `importONNXLayers`.
- TensorFlow-Keras networks, which require Deep Learning Toolbox Importer for TensorFlow-Keras Models support package software. For more information, see `importKerasLayers`.
- Caffe convolutional networks, which require Deep Learning Toolbox Importer for Caffe Models support package software. For more information, see `importCaffeLayers`.

After you import a deep neural network, you can create a policy or value function representation object using a representation object, such as `rlValueRepresentation`.

When you import deep neural network architectures, consider the following.

- Imported architectures must have a single input layer and a single output layer. Therefore, importing entire critic networks with observation and action input layers is not supported.
- The dimensions of the imported network architecture input and output layers must match the dimensions of the corresponding action, observation, or reward dimensions for your environment.
- After importing the network architecture, you must set the names of the input and output layers to match the names of the corresponding action and observation specifications.

For more information on the deep neural network architectures supported for reinforcement learning, see “Create Policy and Value Function Representations” on page 3-2.

Import Actor and Critic for Image Observation Application

As an example, assume that you have an environment with a 50-by-50 grayscale image observation signal and a continuous action space. To train a policy gradient agent, you require the following function approximators, both of which must have a single 50-by-50 image input observation layer and a single scalar output value.

- **Actor** — Selects an action value based on the current observation
- **Critic** — Estimates the expected long-term reward based on the current observation

Also, assume that you have the following network architectures to import:

- A deep neural network architecture for the actor with a 50-by-50 image input layer and a scalar output layer, which is saved in the ONNX format (`criticNetwork.onnx`).
- A deep neural network architecture for the critic with a 50-by-50 image input layer and a scalar output layer, which is saved in the ONNX format (`actorNetwork.onnx`).

To import the critic and actor networks, use the `importONNXLayers` function without specifying an output layer.

```
criticNetwork = importONNXLayers('criticNetwork.onnx');
actorNetwork = importONNXLayers('actorNetwork.onnx');
```

These commands generate a warning, which states that the network is trainable until an output layer is added. When you use an imported network to create an actor or critic representation, Reinforcement Learning Toolbox software automatically adds an output layer for you.

After you import the networks, create the actor and critic function approximator representations. To do so, first obtain the observation and action specifications from the environment.

```
obsInfo = getObservationInfo(env);  
actInfo = getActionInfo(env);
```

Create the critic representation, specifying the name of the input layer of the critic network as the observation name. Since the critic network has a single observation input and a single action output, use a value-function representation.

```
critic = rlValueRepresentation(criticNetwork,obsInfo,...  
    'Observation',{criticNetwork.Layers(1).Name});
```

Create the actor representation, specifying the name of the input layer of the actor network as the observation name and the output layer of the actor network as the observation name. Since the actor network has a single scalar output, use a deterministic actor representation.

```
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,...  
    'Observation',{actorNetwork.Layers(1).Name},...  
    'Action',{actorNetwork.Layers(end).Name});
```

You can then:

- Create an agent using these representations. For more information, see “Reinforcement Learning Agents” on page 4-2.
- Set the actor and critic representation in an existing agent using `setActor` and `setCritic`, respectively.

See Also

More About

- “Create Policy and Value Function Representations” on page 3-2
- “Reinforcement Learning Agents” on page 4-2

Create Agents

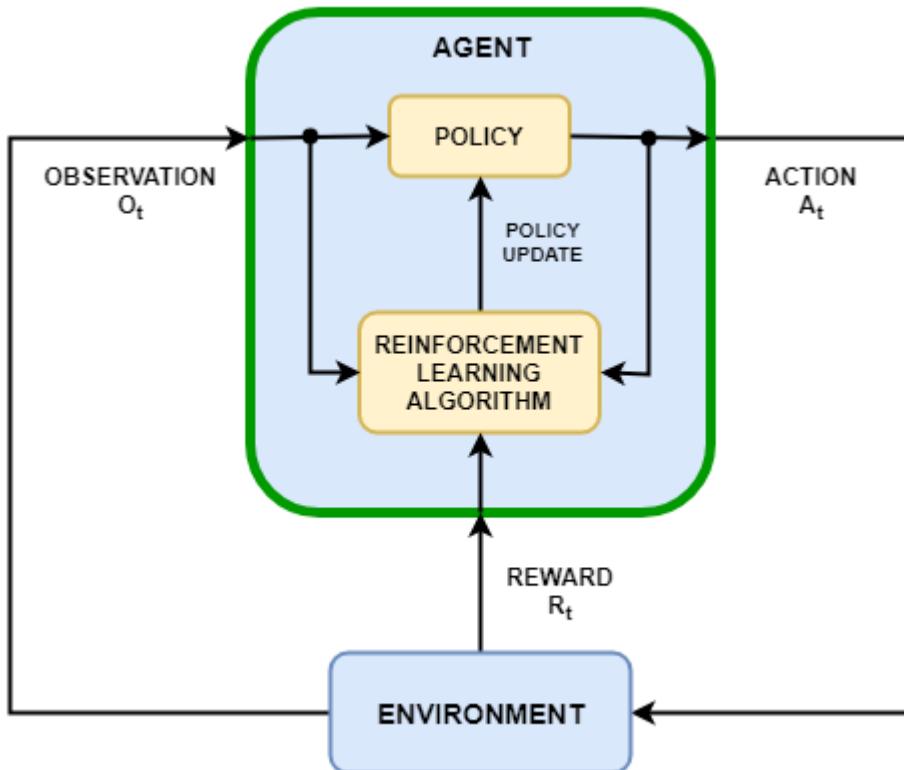
- “Reinforcement Learning Agents” on page 4-2
- “Q-Learning Agents” on page 4-4
- “SARSA Agents” on page 4-6
- “Deep Q-Network Agents” on page 4-8
- “Policy Gradient Agents” on page 4-11
- “Deep Deterministic Policy Gradient Agents” on page 4-14
- “Twin-Delayed Deep Deterministic Policy Gradient Agents” on page 4-17
- “Actor-Critic Agents” on page 4-21
- “Proximal Policy Optimization Agents” on page 4-24
- “Custom Agents” on page 4-28

Reinforcement Learning Agents

The goal of reinforcement learning is to train an agent to complete a task within an uncertain environment. The agent receives observations and a reward from the environment and sends actions to the environment. The reward is a measure of how successful an action is with respect to completing the task goal.

The agent contains two components: a policy and a learning algorithm.

- The policy is a mapping that selects actions based on the observations from the environment. Typically, the policy is a function approximator with tunable parameters, such as a deep neural network.
- The learning algorithm continuously updates the policy parameters based on the actions, observations, and rewards. The goal of the learning algorithm is to find an optimal policy that maximizes the cumulative reward received during the task.



Depending on the learning algorithm, an agent maintains one or more parameterized function approximators for training the policy. There are two types of function approximators.

- **Critics** — For a given observation and action, a critic finds the expected value of the long-term future reward for the task.
- **Actors** — For a given observation, an actor finds the action that maximizes the long-term future reward

For more information on creating actor and critic function approximators, see “Create Policy and Value Function Representations” on page 3-2.

Built-In Agents

Reinforcement Learning Toolbox software provides the following built-in agents. Each agent can be trained in environments with continuous or discrete observation spaces and the following action spaces.

Agent	Actions
"Q-Learning Agents" on page 4-4	Discrete
"SARSA Agents" on page 4-6	Discrete
"Deep Q-Network Agents" on page 4-8	Discrete
"Policy Gradient Agents" on page 4-11	Discrete or continuous
"Deep Deterministic Policy Gradient Agents" on page 4-14	Continuous
"Twin-Delayed Deep Deterministic Policy Gradient Agents" on page 4-17	Continuous
"Actor-Critic Agents" on page 4-21	Discrete or continuous
"Proximal Policy Optimization Agents" on page 4-24	Discrete or continuous

Custom Agents

You can also train policies using other learning algorithms by creating a custom agent. To do so, you create a subclass of a custom agent class, defining the agent behavior using a set of required and optional methods. For more information, see "Custom Agents" on page 4-28.

See Also

[rlACAgent](#) | [rlDDPGAgent](#) | [rlDQNAgent](#) | [rlPGAgent](#) | [rlPPOAgent](#) | [rlQAgent](#) | [rlSARSAAgent](#)

More About

- "What Is Reinforcement Learning?" on page 1-3
- "Train Reinforcement Learning Agents" on page 5-2

Q-Learning Agents

The Q-learning algorithm is a model-free, online, off-policy reinforcement learning method. A Q-learning agent is a value-based reinforcement learning agent which trains a critic to estimate the return or future rewards.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

Q-learning agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Continuous or discrete	Discrete

During training, the agent explores the action space using epsilon-greedy exploration. During each control interval the agent selects a random action with probability ϵ , otherwise it selects an action greedily with respect to the value function with probability $1-\epsilon$. This greedy action is the action for which the value function is greatest.

Critic Function

To estimate the value function, a Q-learning agent maintains a critic $Q(S,A)$, which is a table or function approximator. The critic takes observation S and action A as inputs and outputs the corresponding expectation of the long-term reward.

For more information on creating critics for value function approximation, see “Create Policy and Value Function Representations” on page 3-2.

When training is complete, the trained value function approximator is stored in critic $Q(S,A)$.

Agent Creation

To create a Q-learning agent:

- 1 Create a critic using an `rlQValueRepresentation` object.
- 2 Specify agent options using an `rlQAgentOptions` object.
- 3 Create the agent using an `rlQAgent` object.

Training Algorithm

Q-learning agents use the following training algorithm. To configure the training algorithm, specify options using `rlQAgentOptions`.

- Initialize the critic $Q(S,A)$ with random values.
- For each training episode:
 - 1 Set the initial observation S .
 - 2 Repeat the following for each step of the episode until S is a terminal state:
 - a For the current observation S , select a random action A with probability ϵ . Otherwise, select the action for which the critic value function is greatest.

$$A = \max_A Q(S, A)$$

To specify ϵ and its decay rate, use the `EpsilonGreedyExploration` option.

- b** Execute action A . Observe the reward R and next observation S' .
- c** If S' is a terminal state, set the value function target y to R . Otherwise set it to:

$$y = R + \gamma \max_A Q(S', A)$$

To set the discount factor γ , use the `DiscountFactor` option.

- d** Compute the critic parameter update.

$$\Delta Q = y - Q(S, A)$$

- e** Update the critic using the learning rate α .

$$Q(S, A) = Q(S, A) + \alpha * \Delta Q$$

Specify the learning rate when you create the critic representation by setting the `LearnRate` option in the `rlRepresentationOptions` object.

- f** Set the observation S to S' .

See Also

[rlQAgent](#) | [rlQAgentOptions](#)

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2
- “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8

SARSA Agents

The SARSA algorithm is a model-free, online, on-policy reinforcement learning method. A SARSA agent is a value-based reinforcement learning agent which trains a critic to estimate the return or future rewards.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

SARSA agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Continuous or discrete	Discrete

During training, the agent explores the action space using epsilon-greedy exploration. During each control interval the agent selects a random action with probability ϵ , otherwise it selects an action greedily with respect to the value function with probability $1-\epsilon$. This greedy action is the action for which the value function is greatest.

Critic Function

To estimate the value function, a SARSA agent maintains a critic $Q(S,A)$, which is a table or function approximator. The critic takes observation S and action A as inputs and outputs the corresponding expectation of the long-term reward.

For more information on creating critics for value function approximation, see “Create Policy and Value Function Representations” on page 3-2.

When training is complete, the trained value function approximator is stored in critic $Q(S,A)$.

Agent Creation

To create a SARSA agent:

- 1 Create a critic using an `rlQValueRepresentation` object.
- 2 Specify agent options using an `rlSARSAOptions` object.
- 3 Create the agent using an `rlSARSAAgent` object.

Training Algorithm

SARSA agents use the following training algorithm. To configure the training algorithm, specify options using `rlSARSAOptions`.

- Initialize the critic $Q(S,A)$ with random values.
- For each training episode:
 - 1 Set the initial observation S .
 - 2 For the current observation S , select a random action A with probability ϵ . Otherwise, select the action for which the critic value function is greatest.

$$A = \max_Q(S, A)$$

To specify ϵ and its decay rate, use the `EpsilonGreedyExploration` option.

- 3** Repeat the following for each step of the episode until S is a terminal state:

a Execute action A . Observe the reward R and next observation S' .

b Select an action A' by following the policy from state S' .

$$A' = \max_{A'} Q(S', A')$$

c If S' is a terminal state, set the value function target y to R . Otherwise set it to:

$$y = R + \gamma Q(S', A')$$

To set the discount factor γ , use the `DiscountFactor` option.

- d** Compute the critic parameter update.

$$\Delta Q = y - Q(S, A)$$

- e** Update the critic using the learning rate α .

$$Q(S, A) = Q(S, A) + \alpha * \Delta Q$$

Specify the learning rate when you create the critic representation by setting the `LearnRate` option in the `rlRepresentationOptions` object.

- f** Set the observation S to S' .

- g** Set the action A to A' .

See Also

`rlSARSAAgent` | `rlSARSAAgentOptions`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2
- “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8

Deep Q-Network Agents

The deep Q-network (DQN) algorithm is a model-free, online, off-policy reinforcement learning method. A DQN agent is a value-based reinforcement learning agent that trains a critic to estimate the return or future rewards. DQN is a variant of Q-learning. For more information on Q-learning, see “Q-Learning Agents” on page 4-4.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

DQN agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Continuous or discrete	Discrete

During training, the agent:

- Updates the critic properties at each time step during learning.
- Explores the action space using epsilon-greedy exploration. During each control interval the agent selects a random action with probability ϵ , otherwise it selects an action greedily with respect to the value function with probability $1-\epsilon$. This greedy action is the action for which the value function is greatest.
- Stores past experience using a circular experience buffer. The agent updates the critic based on a mini-batch of experiences randomly sampled from the buffer.

Critic Function

To estimate the value function, a DQN agent maintains two function approximators:

- Critic $Q(S,A)$ — The critic takes observation S and action A as inputs and outputs the corresponding expectation of the long-term reward.
- Target critic $Q'(S,A)$ — To improve the stability of the optimization, the agent periodically updates the target critic based on the latest critic parameter values.

Both $Q(S,A)$ and $Q'(S,A)$ have the same structure and parameterization.

For more information on creating critics for value function approximation, see “Create Policy and Value Function Representations” on page 3-2.

When training is complete, the trained value function approximator is stored in critic $Q(S,A)$.

Agent Creation

To create a DQN agent:

- Create a critic using an `rlQValueRepresentation` object.
- Specify agent options using an `rlDQNAgentOptions` object.
- Create the agent using an `rlDQNAgent` object.

DQN agents support critics that use recurrent deep neural networks as functions approximators.

Training Algorithm

DQN agents use the following training algorithm, in which they update their critic model at each time step. To configure the training algorithm, specify options using `rldqnAgentOptions`.

- Initialize the critic $Q(s,a)$ with random parameter values θ_Q , and initialize the target critic with the same values: $\theta_{Q'} = \theta_Q$.
- For each training time step:
 - 1 For the current observation S , select a random action A with probability ϵ . Otherwise, select the action for which the critic value function is greatest.

$$A = \underset{A}{\operatorname{argmax}} Q(S, A | \theta_Q)$$

To specify ϵ and its decay rate, use the `EpsilonGreedyExploration` option.

- 2 Execute action A . Observe the reward R and next observation S' .
- 3 Store the experience (S, A, R, S') in the experience buffer.
- 4 Sample a random mini-batch of M experiences (S_i, A_i, R_i, S'_i) from the experience buffer. To specify M , use the `MiniBatchSize` option.
- 5 If S'_i is a terminal state, set the value function target y_i to R_i . Otherwise set it to:

$$A_{\max} = \underset{A'}{\operatorname{argmax}} Q(S'_i, A' | \theta_Q) \quad (\text{double DQN})$$

$$y_i = R_i + \gamma Q'(S'_i, A_{\max} | \theta_Q)$$

$$y_i = R_i + \gamma \max_{A'} Q'(S'_i, A' | \theta_{Q'}) \quad (\text{DQN})$$

To set the discount factor γ , use the `DiscountFactor` option. To use double DQN, set the `UseDoubleDQN` option to `true`.

- 6 Update the critic parameters by one-step minimization of the loss L across all sampled experiences.

$$L_k = \frac{1}{M} \sum_{i=1}^M (y_i - Q(S_i, A_i | \theta_Q))^2$$

- 7 Update the target critic parameters depending on the target update method. For more information, see “Target Update Methods” on page 4-9.
- 8 Update the probability threshold ϵ for selecting a random action based on the decay rate specified in the `EpsilonGreedyExploration` option.

Target Update Methods

DQN agents update their target critic parameters using one of the following target update methods.

- **Smoothing** — Update the target parameters at every time step using smoothing factor τ . To specify the smoothing factor, use the `TargetSmoothFactor` option.
- **Periodic** — Update the target parameters periodically without smoothing (`TargetSmoothFactor` = 1). To specify the update period, use the `TargetUpdateFrequency` parameter.

- **Periodic Smoothing** — Update the target parameters periodically with smoothing.

To configure the target update method, create a `rlDQNAgentOptions` object, and set the `TargetUpdateFrequency` and `TargetSmoothFactor` parameters as shown in the following table.

Update Method	TargetUpdateFrequency	TargetSmoothFactor
Smoothing (default)	1	Less than 1
Periodic	Greater than 1	1
Periodic smoothing	Greater than 1	Less than 1

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing Atari With Deep Reinforcement Learning,” *NIPS Deep Learning Workshop*, 2013.

See Also

`rlDQNAgent` | `rlDQNAgentOptions`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2

Policy Gradient Agents

The policy gradient (PG) algorithm is a model-free, online, on-policy reinforcement learning method. A PG agent is a policy-based reinforcement learning agent which directly computes an optimal policy that maximizes the long-term reward.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

PG agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Discrete or continuous	Discrete or continuous

During training, a PG agent:

- Estimates probabilities of taking each action in the action space and randomly selects actions based on the probability distribution.
- Completes a full training episode using the current policy before learning from the experience and updating the policy parameters.

Actor and Critic Functions

PG agents represent the policy using an actor function approximator $\mu(S)$. The actor takes observation S and outputs the probabilities of taking each action in the action space when in state S .

To reduce the variance during gradient estimation, PG agents can use a baseline value function, which is estimated using a critic function approximator, $V(S)$. The critic computes the value function for a given observation state.

For more information on creating actors and critics for function approximation, see “Create Policy and Value Function Representations” on page 3-2.

Agent Creation

To create a PG agent:

- Create an actor representation using an `rlStochasticActorRepresentation` object.
- If you are using a baseline function, create a critic using an `rlValueRepresentation` object.
- Specify agent options using the `rlPAGOptions` object.
- Create the agent using an `rlPAG` object.

Training Algorithm

PG agents use the REINFORCE (Monte-Carlo policy gradient) algorithm either with or without a baseline. To configure the training algorithm, specify options using `rlPAGOptions`.

REINFORCE Algorithm

- Initialize the actor $\mu(S)$ with random parameter values θ_μ .

- 2** For each training episode, generate the episode experience by following actor policy $\mu(S)$. To select an action, the actor generates probabilities for each action in the action space, then the agent randomly selects an action based on the probability distribution. The agent takes actions until it reaches the terminal state, S_T . The episode experience consists of the sequence:

$$S_0, A_0, R_1, S_1, \dots, S_{T-1}, A_{T-1}, R_T, S_T$$

Here, S_t is a state observation, A_{t+1} is an action taken from that state, S_{t+1} is the next state, and R_{t+1} is the reward received for moving from S_t to S_{t+1} .

- 3** For each state in the episode sequence; that is, for $t = 1, 2, \dots, T-1$, calculate the return G_t , which is the discounted future reward.

$$G_t = \sum_{k=t}^T \gamma^{k-t} R_k$$

- 4** Accumulate the gradients for the actor network by following the policy gradient to maximize the expected discounted reward. If the `EntropyLossWeight` option is greater than zero, then additional gradients are accumulated to minimize the entropy loss function.

$$d\theta_\mu = \sum_{t=1}^{T-1} G_t \nabla_{\theta_\mu} \ln \mu(S_t | \theta_\mu)$$

- 5** Update the actor parameters by applying the gradients.

$$\theta_\mu = \theta_\mu + \alpha d\theta_\mu$$

Here, α is the learning rate of the actor. Specify the learning rate when you create the actor representation by setting the `LearnRate` option in the `rlRepresentationOptions` object. For simplicity, this step shows a gradient update using basic stochastic gradient descent. The actual gradient update method depends on the optimizer specified using `rlRepresentationOptions`.

- 6** Repeat steps 2 through 5 for each training episode until training is complete.

REINFORCE with Baseline Algorithm

- 1** Initialize the actor $\mu(S)$ with random parameter values θ_μ .
- 2** Initialize the critic $V(S)$ with random parameter values θ_V .
- 3** For each training episode, generate the episode experience by following actor policy $\mu(S)$. The episode experience consists of the sequence:

$$S_0, A_0, R_1, S_1, \dots, S_{T-1}, A_{T-1}, R_T, S_T$$

- 4** For $t = 1, 2, \dots, T$:

- Calculate the return G_t , which is the discounted future reward.

$$G_t = \sum_{k=t}^T \gamma^{k-t} R_k$$

- Compute the advantage function δ_t using the baseline value function estimate from the critic.

$$\delta_t = G_t - V(S_t | \theta_V)$$

- 5** Accumulate the gradients for the critic network.

$$d\theta_V = \sum_{t=1}^{T-1} \delta_t \nabla_{\theta_V} V(S_t | \theta_V)$$

- 6** Accumulate the gradients for the actor network. If the `EntropyLossWeight` option is greater than zero, then additional gradients are accumulated to minimize the entropy loss function.

$$d\theta_\mu = \sum_{t=1}^{T-1} \delta_t \nabla_{\theta_\mu} \ln \mu(S_t | \theta_\mu)$$

- 7** Update the critic parameters θ_V .

$$\theta_V = \theta_V + \beta d\theta_V$$

Here, β is the learning rate of the critic. Specify the learning rate when you create the critic representation by setting the `LearnRate` option in the `rlRepresentationOptions` object.

- 8** Update the actor parameters θ_μ .

$$\theta_\mu = \theta_\mu + \alpha d\theta_\mu$$

- 9** Repeat steps 3 through 8 for each training episode until training is complete.

For simplicity, this actor and critic updates in this algorithm show a gradient update using basic stochastic gradient descent. The actual gradient update method depends on the optimizer specified using `rlRepresentationOptions`.

References

- [1] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, issue 3-4, pp. 229-256, 1992.

See Also

`rlPAGAgent` | `rlPAGAgentOptions`

More About

- "Reinforcement Learning Agents" on page 4-2
- "Create Policy and Value Function Representations" on page 3-2
- "Train Reinforcement Learning Agents" on page 5-2

Deep Deterministic Policy Gradient Agents

The deep deterministic policy gradient (DDPG) algorithm is a model-free, online, off-policy reinforcement learning method. A DDPG agent is an actor-critic reinforcement learning agent that computes an optimal policy that maximizes the long-term reward.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

DDPG agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Continuous or discrete	Continuous

During training, a DDPG agent:

- Updates the actor and critic properties at each time step during learning.
- Stores past experience using a circular experience buffer. The agent updates the actor and critic using a mini-batch of experiences randomly sampled from the buffer.
- Perturbs the action chosen by the policy using a stochastic noise model at each training step.

Actor and Critic Function

To estimate the policy and value function, a DDPG agent maintains four function approximators:

- Actor $\mu(S)$ — The actor takes observation S and outputs the corresponding action that maximizes the long-term reward.
- Target actor $\mu'(S)$ — To improve the stability of the optimization, the agent periodically updates the target actor based on the latest actor parameter values.
- Critic $Q(S,A)$ — The critic takes observation S and action A as inputs and outputs the corresponding expectation of the long-term reward.
- Target critic $Q'(S,A)$ — To improve the stability of the optimization, the agent periodically updates the target critic based on the latest critic parameter values.

Both $Q(S,A)$ and $Q'(S,A)$ have the same structure and parameterization, and both $\mu(S)$ and $\mu'(S)$ have the same structure and parameterization.

When training is complete, the trained optimal policy is stored in actor $\mu(S)$.

For more information on creating actors and critics for function approximation, see “Create Policy and Value Function Representations” on page 3-2.

Agent Creation

To create a DDPG agent:

- Create an actor using an `rlDeterministicActorRepresentation` object.
- Create a critic using an `rlQValueRepresentation` object.
- Specify agent options using an `rlDDPGAgentOptions` object.

- 4 Create the agent using an `rlDDPGAgent` object.

Training Algorithm

DDPG agents use the following training algorithm, in which they update their actor and critic models at each time step. To configure the training algorithm, specify options using `rlDDPGAgentOptions`.

- Initialize the critic $Q(S, A)$ with random parameter values θ_Q , and initialize the target critic with the same random parameter values: $\theta_{Q'} = \theta_Q$.
- Initialize the actor $\mu(S)$ with random parameter values θ_μ , and initialize the target actor with the same parameter values: $\theta_{\mu'} = \theta_\mu$.
- For each training time step:
 - 1 For the current observation S , select action $A = \mu(S) + N$, where N is stochastic noise from the noise model. To configure the noise model, use the `NoiseOptions` option.
 - 2 Execute action A . Observe the reward R and next observation S' .
 - 3 Store the experience (S, A, R, S') in the experience buffer.
 - 4 Sample a random mini-batch of M experiences (S_i, A_i, R_i, S'_i) from the experience buffer. To specify M , use the `MiniBatchSize` option.
 - 5 If S'_i is a terminal state, set the value function target y_i to R_i . Otherwise set it to:

$$y_i = R_i + \gamma Q'(S'_i, \mu'(S'_i | \theta_\mu) | \theta_{Q'})$$

The value function target is the sum of the experience reward R_i and the discounted future reward. To specify the discount factor γ , use the `DiscountFactor` option.

To compute the cumulative reward, the agent first computes a next action by passing the next observation S'_i from the sampled experience to the target actor. The agent finds the cumulative reward by passing the next action to the target critic.

- 6 Update the critic parameters by minimizing the loss L across all sampled experiences.

$$L_k = \frac{1}{M} \sum_{i=1}^M (y_i - Q(S_i, A_i | \theta_Q))^2$$

- 7 Update the actor parameters using the following sampled policy gradient to maximize the expected discounted reward.

$$\nabla_{\theta_\mu} J \approx \frac{1}{M} \sum_{i=1}^M G_{ai} G_{\mu i}$$

$$G_{ai} = \nabla_A Q(S_i, A | \theta_Q) \quad \text{where } A = \mu(S_i | \theta_\mu)$$

$$G_{\mu i} = \nabla_{\theta_\mu} \mu(S_i | \theta_\mu)$$

Here, G_{ai} is the gradient of the critic output with respect to the action computed by the actor network, and $G_{\mu i}$ is the gradient of the actor output with respect to the actor parameters. Both gradients are evaluated for observation S_i .

- 8 Update the target actor and critic parameters depending on the target update method. For more information see “Target Update Methods” on page 4-16.

For simplicity, the actor and critic updates in this algorithm show a gradient update using basic stochastic gradient descent. The actual gradient update method depends on the optimizer specified using `rlRepresentationOptions`.

Target Update Methods

DDPG agents update their target actor and critic parameters using one of the following target update methods.

- **Smoothing** — Update the target parameters at every time step using smoothing factor τ . To specify the smoothing factor, use the `TargetSmoothFactor` option.

$$\theta_Q' = \tau\theta_Q + (1 - \tau)\theta_Q \quad (\text{critic parameters})$$

$$\theta_\mu' = \tau\theta_\mu + (1 - \tau)\theta_\mu \quad (\text{actor parameters})$$

- **Periodic** — Update the target parameters periodically without smoothing (`TargetSmoothFactor` = 1). To specify the update period, use the `TargetUpdateFrequency` parameter.
- **Periodic Smoothing** — Update the target parameters periodically with smoothing.

To configure the target update method, create a `rlDDPGAgentOptions` object, and set the `TargetUpdateFrequency` and `TargetSmoothFactor` parameters as shown in the following table.

Update Method	TargetUpdateFrequency	TargetSmoothFactor
Smoothing (default)	1	Less than 1
Periodic	Greater than 1	1
Periodic smoothing	Greater than 1	Less than 1

References

- [1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. “Continuous control with deep reinforcement learning,” *International Conference on Learning Representations*, 2016.

See Also

`rlDDPGAgent` | `rlDDPGAgentOptions`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2

Twin-Delayed Deep Deterministic Policy Gradient Agents

The twin-delayed deep deterministic policy gradient (TD3) algorithm is a model-free, online, off-policy reinforcement learning method. A TD3 agent is an actor-critic reinforcement learning agent that computes an optimal policy that maximizes the long-term reward.

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

The TD3 algorithm is an extension of the DDPG algorithm. DDPG agents can overestimate value functions, which can produce suboptimal policies. To reduce value function overestimation includes the following modifications of the DDPG algorithm.

- 1 A TD3 agent learns two Q-value functions and uses the minimum value function estimate during policy updates.
- 2 A TD3 agent updates the policy and targets less frequently than the Q functions.
- 3 When updating the policy, a TD3 agent adds noise to the target action, which makes the policy less likely to exploit actions with high Q-value estimates.

You can use a TD3 agent to implement one of the following training algorithms, depending on the number of critics you specify.

- TD3 — Train the agent with two Q-value functions. This algorithm implements all three of the preceding modifications.
- Delayed DDPG — Train the agent with a single Q-value function. This algorithm trains a DDPG agent with target policy smoothing and delayed policy and target updates.

TD3 agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Continuous or discrete	Continuous

During training, a TD3 agent:

- Updates the actor and critic properties at each time step during learning.
- Stores past experience using a circular experience buffer. The agent updates the actor and critic using a mini-batch of experiences randomly sampled from the buffer.
- Perturbs the action chosen by the policy using a stochastic noise model at each training step.

Actor and Critic Function

To estimate the policy and value function, a TD3 agent maintains the following function approximators:

- Deterministic actor $\mu(S)$ — The actor takes observation S and outputs the corresponding action that maximizes the long-term reward.
- Target actor $\mu'(S)$ — To improve the stability of the optimization, the agent periodically updates the target actor based on the latest actor parameter values.
- One or two Q-value critics $Q_k(S, A)$ — The critics take observation S and action A as inputs and output the corresponding expectation of the long-term reward.

- One or two target critics $Q'_k(S,A)$ — To improve the stability of the optimization, the agent periodically updates the target critics based on the latest parameter values of the critics. The number of target critics matches the number of critics.

Both $\mu(S)$ and $\mu'(S)$ have the same structure and parameterization.

For each critic, $Q_k(S,A)$ and $Q'_k(S,A)$ have the same structure and parameterization.

When using two critics, $Q_1(S,A)$ and $Q_2(S,A)$, each critic can have a different structure, though TD3 works best when the critics have the same structure. When the critics have the same structure, they must have different initial parameter values.

When training is complete, the trained optimal policy is stored in actor $\mu(S)$.

For more information on creating actors and critics for function approximation, see “Create Policy and Value Function Representations” on page 3-2.

Agent Creation

To create a TD3 agent:

- 1 Create an actor using an `rlDeterministicActorRepresentation` object.
- 2 Create one or two critics using `rlQValueRepresentation` objects.
- 3 Specify agent options using an `rlTD3AgentOptions` object.
- 4 Create the agent using an `rlTD3Agent` object.

Training Algorithm

TD3 agents use the following training algorithm, in which they update their actor and critic models at each time step. To configure the training algorithm, specify options using `rlDDPGAgentOptions`. Here, $K = 2$ is the number of critics and k is the critic index.

- Initialize each critic $Q_k(S,A)$ with random parameter values θ_{Qk} , and initialize each target critic with the same random parameter values: $\theta_{QK} = \theta_{Qk}$.
- Initialize the actor $\mu(S)$ with random parameter values θ_μ , and initialize the target actor with the same parameter values: $\theta_{\mu'} = \theta_\mu$.
- For each training time step:
 - 1 For the current observation S , select action $A = \mu(S) + N$, where N is stochastic noise from the noise model. To configure the noise model, use the `ExplorationModel` option.
 - 2 Execute action A . Observe the reward R and next observation S' .
 - 3 Store the experience (S,A,R,S') in the experience buffer.
 - 4 Sample a random mini-batch of M experiences (S_i,A_i,R_i,S'_i) from the experience buffer. To specify M , use the `MiniBatchSize` option.
 - 5 If S'_i is a terminal state, set the value function target y_i to R_i . Otherwise set it to:

$$y_i = R_i + \gamma * \min_k(Q'_k(S'_i, \text{clip}(\mu'(S'_i | \theta_\mu) + \varepsilon) | \theta_{Qk}))$$

The value function target is the sum of the experience reward R_i and the minimum discounted future reward from the critics. To specify the discount factor γ , use the `DiscountFactor` option.

To compute the cumulative reward, the agent first computes a next action by passing the next observation S'_i from the sampled experience to the target actor. Then, the agent adds noise ε to the computed action using the `TargetPolicySmoothModel`, and clips the action based on the upper and lower noise limits. The agent finds the cumulative rewards by passing the next action to the target critics.

- 6 At every time training step, update the parameters of each critic by minimizing the loss L_k across all sampled experiences.

$$L_k = \frac{1}{M} \sum_{i=1}^M (y_i - Q(S_i, A_i | \theta_Q))^2$$

- 7 Every D_1 steps, update the actor parameters using the following sampled policy gradient to maximize the expected discounted reward. To set D_1 , use the `PolicyUpdateFrequency` option.

$$\nabla_{\theta_\mu} J \approx \frac{1}{M} \sum_{i=1}^M G_{ai} G_{\mu i}$$

$$G_{ai} = \nabla_A \min_k(Q_k(S_i, A | \theta_Q)) \quad \text{where } A = \mu(S_i | \theta_\mu)$$

$$G_{\mu i} = \nabla_{\theta_\mu} \mu(S_i | \theta_\mu)$$

Here, G_{ai} is the gradient of the minimum critic output with respect to the action computed by the actor network, and $G_{\mu i}$ is the gradient of the actor output with respect to the actor parameters. Both gradients are evaluated for observation S_i .

- 8 Every D_2 steps, update the target actor and critics depending on the target update method. To specify D_2 , use the `TargetUpdateFrequency` option. For more information, see “Target Update Methods” on page 4-19.

For simplicity, the actor and critic updates in this algorithm show a gradient update using basic stochastic gradient descent. The actual gradient update method depends on the optimizer specified using `rlRepresentationOptions`.

Target Update Methods

TD3 agents update their target actor and critic parameters using one of the following target update methods.

- **Smoothing** — Update the target parameters at every time step using smoothing factor τ . To specify the smoothing factor, use the `TargetSmoothFactor` option.
- $$\theta_{Qk} = \tau \theta_{Qk} + (1 - \tau) \theta_{Q'k} \quad (\text{critic parameters})$$
- $$\theta_{\mu'} = \tau \theta_\mu + (1 - \tau) \theta_{\mu'} \quad (\text{actor parameters})$$
- **Periodic** — Update the target parameters periodically without smoothing (`TargetSmoothFactor` = 1). To specify the update period, use the `TargetUpdateFrequency` parameter.

$$\theta_{Qk} = \theta_{Qk}$$

$$\theta_{\mu'} = \theta_\mu$$

- **Periodic Smoothing** — Update the target parameters periodically with smoothing.

To configure the target update method, create a `rLTD3AgentOptions` object, and set the `TargetUpdateFrequency` and `TargetSmoothFactor` parameters as shown in the following table.

Update Method	TargetUpdateFrequency	TargetSmoothFactor
Smoothing (default)	1	Less than 1
Periodic	Greater than 1	1
Periodic smoothing	Greater than 1	Less than 1

References

[1] Fujimoto, Scott, Herke van Hoof, and David Meger. 'Addressing Function Approximation Error in Actor-Critic Methods'. *ArXiv:1802.09477 [Cs, Stat]*, 22 October 2018. <https://arxiv.org/abs/1802.09477>.

See Also

`rLTD3Agent` | `rLTD3AgentOptions`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2
- “Train Biped Robot to Walk Using Reinforcement Learning Agents” on page 5-134

Actor-Critic Agents

You can use the actor-critic (AC) agent, which uses a model-free, online, on-policy reinforcement learning method, to implement actor-critic algorithms, such as A2C and A3C. The goal of this agent is to optimize the policy (actor) directly and train a critic to estimate the return or future rewards. [1]

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

AC agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Discrete or continuous	Discrete or continuous

During training, an AC agent:

- Estimates probabilities of taking each action in the action space and randomly selects actions based on the probability distribution.
- Interacts with the environment for multiple steps using the current policy before updating the actor and critic properties.

Actor and Critic Function

To estimate the policy and value function, an AC agent maintains two function approximators:

- Actor $\mu(S)$ — The actor takes observation S and outputs the probabilities of taking each action in the action space when in state S .
- Critic $V(S)$ — The critic takes observation S and outputs the corresponding expectation of the discounted long-term reward.

When training is complete, the trained optimal policy is stored in actor $\mu(S)$.

For more information on creating actors and critics for function approximation, see “Create Policy and Value Function Representations” on page 3-2.

Agent Creation

To create an AC agent:

- Create an actor using an `rlStochasticActorRepresentation` object.
- Create a critic using an `rlValueRepresentation` object.
- Specify agent options using an `rlACAgentOptions` object.
- Create the agent using an `rlACAgent` object.

Training Algorithm

AC agents use the following training algorithm. To configure the training algorithm, specify options using an `rlACAgentOptions` object.

- Initialize the actor $\mu(S)$ with random parameter values θ_μ .

- 2 Initialize the critic $V(S)$ with random parameter values θ_V .
- 3 Generate N experiences by following the current policy. The episode experience sequence is:

$$S_{ts}, A_{ts}, R_{ts+1}, S_{ts+1}, \dots, S_{ts+N-1}, A_{ts+N-1}, R_{ts+N}, S_{ts+N}$$

Here, S_t is a state observation, A_t is an action taken from that state, S_{t+1} is the next state, and R_{t+1} is the reward received for moving from S_t to S_{t+1} .

When in state S_t , the agent computes the probability of taking each action in the action space using $\mu(S_t)$ and randomly selects action A_t based on the probability distribution.

ts is the starting time step of the current set of N experiences. At the beginning of the training episode, $ts = 1$. For each subsequent set of N experiences in the same training episode, $ts = ts + N$.

For each training episode that does not contain a terminal state, N is equal to the `NumStepsToLookAhead` option value. Otherwise, N is less than `NumStepsToLookAhead` and S_N is the terminal state.

- 4 For each episode step $t = ts+1, ts+2, \dots, ts+N$, compute the return G_t , which is the sum of the reward for that step and the discounted future reward. If S_{ts+N} is not a terminal state, the discounted future reward includes the discounted state value function, computed using the critic network V .

$$G_t = \sum_{k=t}^{ts+N} (\gamma^k - t R_k) + b \gamma^{N-t+1} V(S_{ts+N} | \theta_V)$$

Here, b is 0 if S_{ts+N} is a terminal state and 1 otherwise.

To specify the discount factor γ , use the `DiscountFactor` option.

- 5 Compute the advantage function D_t .

$$D_t = G_t - V(S_t | \theta_V)$$

- 6 Accumulate the gradients for the actor network by following the policy gradient to maximize the expected discounted reward.

$$d\theta_\mu = \sum_{t=1}^N \nabla_{\theta_\mu} \ln \mu(S_t | \theta_\mu) * D_t$$

- 7 Accumulate the gradients for the critic network by minimizing the mean square error loss between the estimated value function $V(t)$ and the computed target return G_t across all N experiences. If the `EntropyLossWeight` option is greater than zero, then additional gradients are accumulated to minimize the entropy loss function.

$$d\theta_V = \sum_{t=1}^N \nabla_{\theta_V} (G_t - V(S_t | \theta_V))^2$$

- 8 Update the actor parameters by applying the gradients.

$$\theta_\mu = \theta_\mu + \alpha d\theta_\mu$$

Here, α is the learning rate of the actor. Specify the learning rate when you create the actor representation by setting the `LearnRate` option in the `rlRepresentationOptions` object.

-
- 9 Update the critic parameters by applying the gradients.

$$\theta_V = \theta_V + \beta d\theta_V$$

Here, β is the learning rate of the critic. Specify the learning rate when you create the critic representation by setting the `LearnRate` option in the `rlRepresentationOptions` object.

- 10 Repeat steps 3 through 9 for each training episode until training is complete.

For simplicity, the actor and critic updates in this algorithm show a gradient update using basic stochastic gradient descent. The actual gradient update method depends on the optimizer specified using `rlRepresentationOptions`.

References

- [1] Mnih, V, et al. "Asynchronous methods for deep reinforcement learning," *International Conference on Machine Learning*, 2016.

See Also

`rlACAgent` | `rlACAgentOptions`

More About

- "Reinforcement Learning Agents" on page 4-2
- "Create Policy and Value Function Representations" on page 3-2
- "Train Reinforcement Learning Agents" on page 5-2

Proximal Policy Optimization Agents

The proximal policy optimization (PPO) is a model-free, online, on-policy, policy gradient reinforcement learning method. This algorithm is a type of policy gradient training that alternates between sampling data through environmental interaction and optimizing a clipped surrogate objective function using stochastic gradient descent. The clipped surrogate objective function improves training stability by limiting the size of the policy change at each step. [1]

For more information on the different types of reinforcement learning agents, see “Reinforcement Learning Agents” on page 4-2.

PPO agents can be trained in environments with the following observation and action spaces.

Observation Space	Action Space
Discrete or continuous	Discrete or continuous

During training, a PPO agent:

- Estimates probabilities of taking each action in the action space and randomly selects actions based on the probability distribution.
- Interacts with the environment for multiple steps using the current policy before using mini-batches to update the actor and critic properties over multiple epochs.

Actor and Critic Function

To estimate the policy and value function, a PPO agent maintains two function approximators:

- Actor $\mu(S)$ — The actor takes observation S and outputs the probabilities of taking each action in the action space when in state S .
- Critic $V(S)$ — The critic takes observation S and outputs the corresponding expectation of the discounted long-term reward.

When training is complete, the trained optimal policy is stored in actor $\mu(S)$.

For more information on creating actors and critics for function approximation, see “Create Policy and Value Function Representations” on page 3-2.

Agent Creation

To create a PPO agent:

- Create an actor using an `rlStochasticActorRepresentation` object.
- Create a critic using an `rlValueRepresentation` object.
- Specify agent options using an `rlPPOAgentOptions` object.
- Create the agent using the `rlPPOAgent` function.

PPO agents support actors and critics that use recurrent deep neural networks as functions approximators.

Training Algorithm

PPO agents use the following training algorithm. To configure the training algorithm, specify options using an `rlPPOAgentOptions`.

- 1 Initialize the actor $\mu(S)$ with random parameter values θ_μ .
- 2 Initialize the critic $V(S)$ with random parameter values θ_V .
- 3 Generate N experiences by following the current policy. The experience sequence is:

$$S_{ts}, A_{ts}, R_{ts+1}, S_{ts+1}, \dots, S_{ts+N-1}, A_{ts+N-1}, R_{ts+N}, S_{ts+N}$$

Here, S_t is a state observation, A_t is an action taken from that state, S_{t+1} is the next state, and R_{t+1} is the reward received for moving from S_t to S_{t+1} .

When in state S_t , the agent computes the probability of taking each action in the action space using $\mu(S_t)$ and randomly selects action A_t based on the probability distribution.

ts is the starting time step of the current set of N experiences. At the beginning of the training episode, $ts = 1$. For each subsequent set of N experiences in the same training episode, $ts \leftarrow ts + N$.

For each experience sequence that does not contain a terminal state, N is equal to the `ExperienceHorizon` option value. Otherwise, N is less than `ExperienceHorizon` and S_N is the terminal state.

- 4 For each episode step $t = ts+1, ts+2, \dots, ts+N$, compute the return and advantage function using the method specified by the `AdvantageEstimateMethod` option.
 - **Finite Horizon** (`AdvantageEstimateMethod = "finite-horizon"`) — Compute the return G_t , which is the sum of the reward for that step and the discounted future reward. [2]

$$G_t = \sum_{k=t}^{ts+N} (\gamma^k - t R_k) + b \gamma^{N-t+1} V(S_{ts+N} | \theta_V)$$

Here, b is 0 if S_{ts+N} is a terminal state and 1 otherwise. That is, if S_{ts+N} is not a terminal state, the discounted future reward includes the discounted state value function, computed using the critic network V .

Compute the advantage function D_t .

$$D_t = G_t - V(S_t | \theta_V)$$

- **Generalized Advantage Estimator** (`AdvantageEstimateMethod = "gae"`) — Compute the advantage function D_t , which is the discounted sum of temporal difference errors. [3]

$$D_t = \sum_{k=t}^{ts+N-1} (\gamma \lambda)^{k-t} \delta_k$$

$$\delta_k = R_t + b \gamma V(S_t | \theta_V)$$

Here, b is 0 if S_{ts+N} is a terminal state and 1 otherwise. λ is a smoothing factor specified using the `GAEFactor` option.

Compute the return G_t .

$$G_t = D_t - V(S_t|\theta_V)$$

To specify the discount factor γ for either method, use the `DiscountFactor` option.

- 5 Learn from experience mini-batches over K epochs. To specify K , use the `NumEpoch` option. For each learning epoch:
 - a Sample a random mini-batch data set of size M from the current set of experience. To specify M , use the `MiniBatchSize` option. Each element of the mini-batch data set contains a current experience and the corresponding return and advantage function values.
 - b Update the critic parameters by minimizing the loss L_{critic} across all sampled mini-batch data.

$$L_{critic}(\theta_V) = \frac{1}{M} \sum_{i=1}^M (G_i - V(S_i|\theta_V))^2$$

- c Update the actor parameters by minimizing the loss L_{actor} across all sampled mini-batch data. If the `EntropyLossWeight` option is greater than zero, then additional entropy loss is added to L_{actor} , which encourages policy exploration.

$$L_{actor}(\theta_\mu) = -\frac{1}{M} \sum_{i=1}^M \min(r_i(\theta_\mu) * D_i, c_i(\theta_\mu) * D_i)$$

$$r_i(\theta_\mu) = \frac{\mu_{Ai}(S_i|\theta_\mu)}{\mu_{Ai}(S_i|\theta_{\mu, old})}$$

$$c_i(\theta_\mu) = \max(\min(r_i(\theta_\mu), 1 + \varepsilon), 1 - \varepsilon)$$

Here:

- D_i, G_i are the advantage function and return value for the i th element of the mini-batch, respectively.
- $\mu_i(S_i|\theta_\mu)$ is the probability of taking action A_i when in state S_i , given the updated policy parameters θ_μ .
- $\mu_i(S_i|\theta_{\mu, old})$ is the probability of taking action A_i when in state S_i , given the previous policy parameters ($\theta_{\mu, old}$) from before the current learning epoch.
- ε is the clip factor specified using the `ClipFactor` option.

- 6 Repeat steps 3 through 5 until the training episode reaches a terminal state.

References

- [1] Schulman, J., et al. "Proximal Policy Optimization Algorithms," Technical Report, *ArXiv*, 2017.
- [2] Mnih, V., et al. "Asynchronous methods for deep reinforcement learning," *International Conference on Machine Learning*, 2016.
- [3] Schulman, J., et al. "High-Dimensional Continuous Control Using Generalized Advantage Estimation," Technical Report, *ArXiv*, 2018.

See Also

[rlPPOAgent](#) | [rlPPOAgentOptions](#)

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2

Custom Agents

To implement your own custom reinforcement learning algorithms, you can create a custom agent by creating a subclass of a custom agent class. For more information about creating MATLAB classes, see “User-Defined Classes” (MATLAB).

Create Template Class

To define your custom agent, first create a class that is a subclass of the `rl.agent.CustomAgent` class. As an example, this topic describes the custom LQR agent trained in “Train Custom LQR Agent” on page 5-153. As a starting point for your own agent, you can open and modify this custom agent class. To add the example files to the MATLAB path and open the file, at the MATLAB command line, type the following code.

```
addpath(fullfile(matlabroot,'examples','rl','main'));
edit LQRCustomAgent.m
```

After saving the class to your own working folder, you can remove the example files from the path.

```
rmpath(fullfile(matlabroot,'examples','rl','main'));
```

This class has the following class definition, which indicates the agent class name and the associated abstract agent.

```
classdef LQRCustomAgent < rl.agent.CustomAgent
```

To define your agent you must specify the following:

- Agent properties
- Constructor function
- Critic representation that estimates the discounted long-term reward (if required for learning)
- Actor representation that selects an action based on the current observation (if required for learning)
- Required agent methods
- Optional agent methods

Agent Properties

In the `properties` section of the class file, specify any parameters necessary for creating and training the agent. These parameters can include:

- Discount factor for discounting future rewards
- Configuration parameters for exploration models, such as noise models or epsilon-greedy exploration
- Experience buffers for using replay memory
- Mini-batch sizes for sampling from the experience buffer
- Number of steps to look ahead during training

For more information on potential agent properties, see the option sets for the built-in Reinforcement Learning Toolbox agents.

The `rl.Agent.CustomAgent` class already includes properties for the agent sample time (`SampleTime`) and the action and observation specifications (`ActionInfo` and `ObservationInfo`, respectively).

The custom LQR agent defines the following agent properties.

```
properties
    % Q
    Q

    % R
    R

    % Feedback gain
    K

    % Discount Factor
    Gamma = 0.95

    % Critic
    Critic

    % Buffer for K
    KBuffer
    % Number of updates for K
    KUpdate = 1

    % Number for estimator update
    EstimateNum = 10
end

properties (Access = private)
    Counter = 1
    YBuffer
    HBuffer
end
```

Constructor Function

To create your custom agent, you must define a constructor function that:

- Defines the action and observation specifications. For more information about creating these specifications, see `rlNumericSpec` and `rlFiniteSetSpec`.
- Creates actor and critic representations as required by your training algorithm. For more information, see “Create Policy and Value Function Representations” on page 3-2.
- Configures agent properties.
- Calls the constructor of the base abstract class.

For example, the `LQRCustomAgent` constructor defines continuous action and observation spaces and creates a critic representation. The `createCritic` function is an optional helper function used for defining the critic representation.

```
function obj = LQRCustomAgent(Q,R,InitialK)
    % Check the number of input arguments
    narginchk(3,3);
```

```
% Call the abstract class constructor
obj = obj@rl.agent.CustomAgent();

% Set the Q and R matrices
obj.Q = Q;
obj.R = R;

% Define the observation and action spaces
obj.ObservationInfo = rlNumericSpec([size(Q,1),1]);
obj.ActionInfo = rlNumericSpec([size(R,1),1]);

% Create the critic representation
obj.Critic = createCritic(obj);

% Initialize the gain matrix
obj.K = InitialK;

% Initialize the experience buffers
obj.YBuffer = zeros(obj.EstimateNum,1);
num = size(Q,1) + size(R,1);
obj.HBuffer = zeros(obj.EstimateNum,0.5*num*(num+1));
obj.KBuffer = cell(1,1000);
obj.KBuffer{1} = obj.K;
end
end
```

Actor and Critic Representations

If your learning algorithm uses a critic representation to estimate the long term reward, an actor for selecting an action, or both, you must add these as agent properties. You must then create these representations when you create your agent; that is, in the constructor function. For more information on creating actors and critics, see “Create Policy and Value Function Representations” on page 3-2.

For example, the custom LQR agent uses a critic representation, stored in its `Critic` property, and no actor. The critic creation is implemented in the `createCritic` helper function, which is called from the `LQRCustomAgent` constructor.

```
function critic = createCritic(obj)
nQ = size(obj.Q,1);
nR = size(obj.R,1);
n = nQ+nR;
w0 = 0.1*ones(0.5*(n+1)*n,1);
critic = rlQValueRepresentation({@(x,u) computeQuadraticBasis(x,u,n),w0},...
    getObservationInfo(obj),getActionInfo(obj));
critic.Options.GradientThreshold = 1;
end
```

In this case, the critic is an `rlQValueRepresentation` object. To create such a representation, you must specify the handle to a custom basis function, in this case the `computeQuadraticBasis` function. For more information on this critic representation, see “Train Custom LQR Agent” on page 5-153.

Required Functions

To create a custom reinforcement learning agent you must define the following implementation functions. To call these functions in your own code, use the wrapper methods from the abstract base

class. For example, to call `getActionImpl`, use `getAction`. The wrapper methods have the same input and output arguments as the implementation methods.

Function	Description
<code>getActionImpl</code>	Selects an action by evaluating the agent policy for a given observation.
<code>getActionWithExplorationImpl</code>	Selects an action using the exploration model of the agent
<code>learnImpl</code>	Learns from the current experiences and returns an action with exploration

Within your implementation functions, to evaluate your actor and critic representations, you can use the `getValue`, `getAction`, and `getMaxQValue` functions. To evaluate:

- An `rlValueRepresentation` critic with only observation input signals, obtain the state value function V using:

```
V = getValue(Critic,Observation);
```

- An `rlQValueRepresentation` critic with both observation and action input signals, obtain the state-action value function Q using:

```
Q = getValue(Critic,[Observation,Action]);
```

- An `rlQValueRepresentation` critic with only observation input signals, obtain the state-action value function Q for all possible discrete actions using:

```
Q = getValue(Critic,Observation);
```

- A discrete action space `rlQValueRepresentation` critic, obtain the maximum Q state-action value function Q for all possible discrete actions using:

```
[MaxQ,MaxActionIndex] = getMaxQValue(Critic,Observation);
```

- An actor representation (`rlStochasticActorRepresentation` or `rlDeterministicActorRepresentation`), obtain the action A using:

```
A = getAction(Actor,Observation);
```

For each of these cases, if your actor or critic network uses a recurrent neural network, the functions can also return the current values of the network state after obtain the corresponding network output.

getActionImpl Function

The `getActionImpl` function is used to evaluate the policy of your agent and select an action. This function must have the following signature, where `obj` is the agent object, `Observation` is the current observation, and `action` is the selected action.

```
function action = getActionImpl(obj,Observation)
```

For the custom LQR agent, you select an action by applying the $u=-Kx$ control law.

```
function action = getActionImpl(obj,Observation)
    % Given the current state of the system, return an action.
    action = -obj.K*Observation{,:};
end
```

getActionWithExplorationImpl Function

The `getActionWithExplorationImpl` function selects an action using the exploration model of your agent. Using this function you can implement algorithms such as epsilon-greedy exploration. This function must have the following signature, where `obj` is the agent object, `Observation` is the current observation, and `action` is the selected action.

```
function action = getActionWithExplorationImpl(obj,Observation)
```

For the custom LQR agent, the `getActionWithExplorationImpl` function adds random white noise to an action selected using the current agent policy.

```
function action = getActionWithExplorationImpl(obj,Observation)
    % Given the current observation, select an action
    action = getAction(obj,Observation);

    % Add random noise to the action
    num = size(obj.R,1);
    action = action + 0.1*randn(num,1);
end
```

learnImpl Function

The `learnImpl` function defines how the agent learns from the current experience. This function implements the custom learning algorithm of your agent by updating the policy parameters and selecting an action with exploration. This function must have the following signature, where `obj` is the agent object, `exp` is the current agent experience, and `action` is the selected action.

```
function action = learnImpl(obj,exp)
```

The agent experience, is the cell array `exp = {state,action,reward,nextstate,isdone}`. Here:

- `state` is the current observation.
- `action` is the current action.
- `reward` is the current reward.
- `nextState` is the next observation.
- `isDone` is a logical flag indicating that the training episode is complete.

For the custom LQR agent, the critic parameters are updated every `N` steps.

```
function action = learnImpl(obj,exp)
    % Parse the experience input
    x = exp{1}{1};
    u = exp{2}{1};
    dx = exp{4}{1};
    y = (x'*obj.Q*x + u'*obj.R*u);
    num = size(obj.Q,1) + size(obj.R,1);

    % Wait N steps before updating critic parameters
    N = obj.EstimateNum;
    h1 = computeQuadraticBasis(x,u,num);
    h2 = computeQuadraticBasis(dx,-obj.K*dx,num);
    H = h1 - obj.Gamma* h2;
    if obj.Counter<=N
        obj.YBuffer(obj.Counter) = y;
```

```

obj.HBuffer(obj.Counter,:) = H;
obj.Counter = obj.Counter + 1;
else
    % Update the critic parameters based on the batch of
    % experiences
    H_buf = obj.HBuffer;
    y_buf = obj.YBuffer;
    theta = (H_buf'*H_buf)\H_buf'*y_buf;
    obj.Critic = setLearnableParameters(obj.Critic,{theta});

    % Derive a new gain matrix based on the new critic parameters
    obj.K = getNewK(obj);

    % Reset the experience buffers
    obj.Counter = 1;
    obj.YBuffer = zeros(N,1);
    obj.HBuffer = zeros(N,0.5*num*(num+1));
    obj.KUpdate = obj.KUpdate + 1;
    obj.KBuffer{obj.KUpdate} = obj.K;
end

% Find and return an action with exploration
action = getActionWithExploration(obj,exp{4});
end

```

Optional Functions

Optionally, you can define how your agent is reset at the start of training by specifying a `resetImpl` function with the following function signature, where `obj` is the agent object. Using this function, you can set the agent into a know or random condition before training.

```
function resetImpl(ob)
```

Also, you can define any other helper functions in your custom agent class as required. For example, the custom LQR agent defines a `createCritic` function for creating the critic representation and a `getNewK` function that derives the feedback gain matrix from the trained critic parameters.

Create Custom Agent

Once you have defined your custom agent class, create an instance of it in the MATLAB workspace. For example, to create the custom LQR agent, define the `Q`, `R`, `InitialK` values, and call the constructor function.

```

Q = [10,3,1;3,5,4;1,4,9];
R = 0.5*eye(3);
K0 = place(A,B,[0.4,0.8,0.5]);
agent = LQRCustomAgent(Q,R,K0);

```

After validating the environment object, you can use it to train a reinforcement learning agent. For an example that trains the custom LQR agent, see “Train Custom LQR Agent” on page 5-153.

See Also

`train`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Create Policy and Value Function Representations” on page 3-2
- “Train Reinforcement Learning Agents” on page 5-2

Train and Validate Agents

- “Train Reinforcement Learning Agents” on page 5-2
- “Train DQN Agent to Balance Cart-Pole System” on page 5-8
- “Train PG Agent to Balance Cart-Pole System” on page 5-14
- “Train AC Agent to Balance Cart-Pole System” on page 5-19
- “Train PG Agent with Baseline to Control Double Integrator System” on page 5-25
- “Train DDPG Agent to Control Double Integrator System” on page 5-31
- “Train DQN Agent to Swing Up and Balance Pendulum” on page 5-37
- “Train DDPG Agent to Swing Up and Balance Pendulum” on page 5-43
- “Train DDPG Agent to Swing Up and Balance Cart-Pole System” on page 5-50
- “Train DDPG Agent to Swing Up and Balance Pendulum with Bus Signal” on page 5-57
- “Train DDPG Agent to Swing Up and Balance Pendulum with Image Observation” on page 5-64
- “Create Agent Using Deep Network Designer and Train Using Image Observations” on page 5-72
- “Train AC Agent to Balance Cart-Pole System Using Parallel Computing” on page 5-84
- “Train DDPG Agent to Control Flying Robot” on page 5-89
- “Train PPO Agent to Land Rocket” on page 5-95
- “Train DDPG Agent for Adaptive Cruise Control” on page 5-102
- “Train DQN Agent for Lane Keeping Assist” on page 5-110
- “Train DDPG Agent for Path Following Control” on page 5-118
- “Train DQN Agent for Lane Keeping Assist Using Parallel Computing” on page 5-126
- “Train Biped Robot to Walk Using Reinforcement Learning Agents” on page 5-134
- “Quadruped Robot Locomotion Using DDPG Agent” on page 5-145
- “Train Custom LQR Agent” on page 5-153
- “Imitate MPC Controller for Lane Keep Assist” on page 5-157
- “Train DDPG Agent with Pretrained Actor Network” on page 5-165
- “Imitate Nonlinear MPC Controller for Flying Robot” on page 5-174
- “Train Reinforcement Learning Policy Using Custom Training Loop” on page 5-182

Train Reinforcement Learning Agents

Once you have created an environment and reinforcement learning agent, you can train the agent in the environment using the `train` function. To configure your training, use the `rlTrainingOptions` function. For example, create a training option set `opt`, and train agent `agent` in environment `env`.

```
opt = rlTrainingOptions(...  
    'MaxEpisodes',1000,...  
    'MaxStepsPerEpisode',1000,...  
    'StopTrainingCriteria','AverageReward',...  
    'StopTrainingValue',480);  
trainStats = train(agent,env,opt);
```

For more information on creating agents, see “Reinforcement Learning Agents” on page 4-2. For more information on creating environments, see “Create MATLAB Environments for Reinforcement Learning” on page 2-2 and “Create Simulink Environments for Reinforcement Learning” on page 2-5.

`train` updates the agent as training progresses. To preserve the original agent parameters for later use, save the agent to a MAT-file.

```
save("initialAgent.mat","agent")
```

Training terminates automatically when the conditions you specify in the `StopTrainingCriteria` and `StopTrainingValue` options of your `rlTrainingOptions` object are satisfied. To manually terminate training in progress, type **Ctrl+C** or, in the Reinforcement Learning Episode Manager, click **Stop Training**. Because `train` updates the agent at each episode, you can resume training by calling `train(agent,env,trainOpts)` again, without losing the trained parameters learned during the first call to `train`.

Training Algorithm

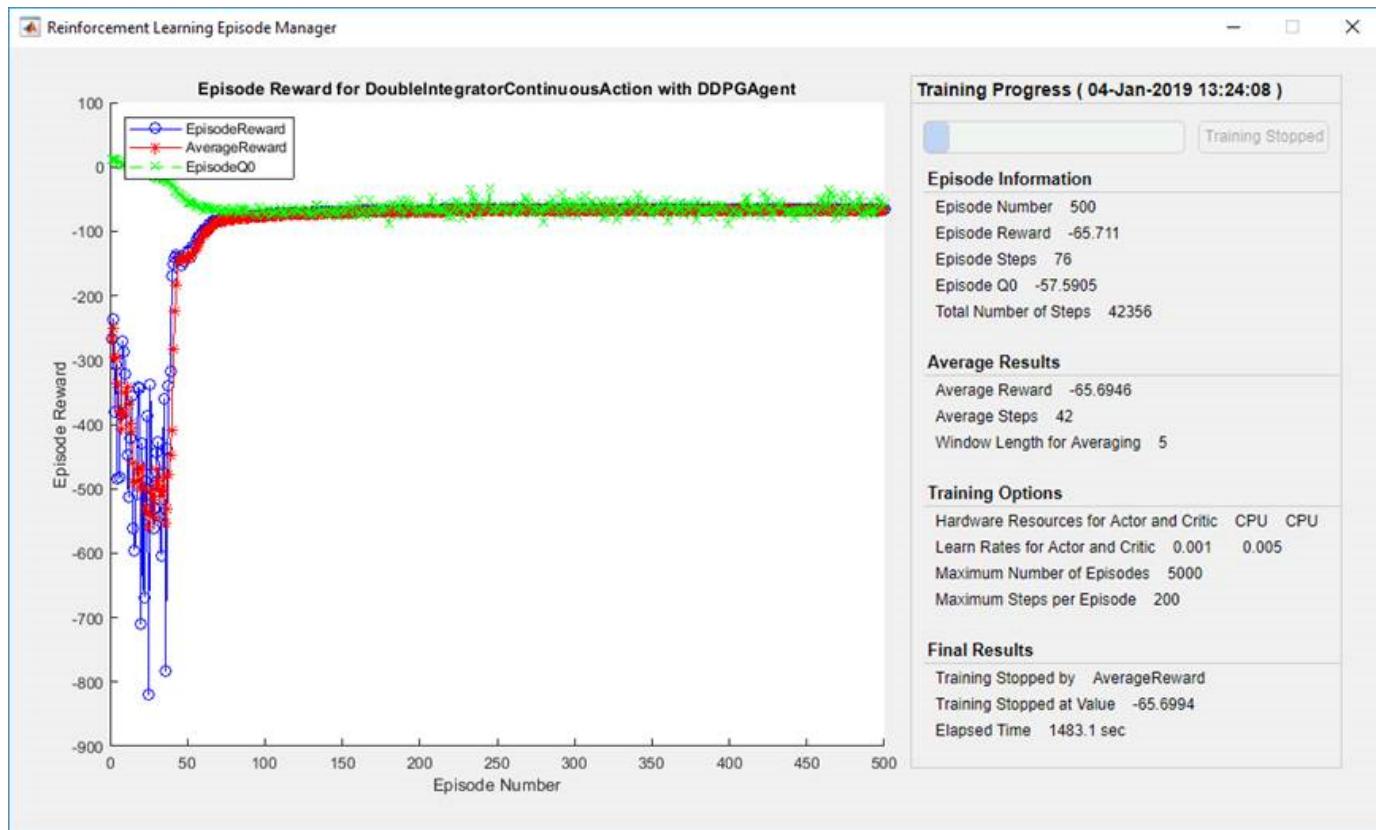
In general, training performs the following steps.

- 1 Initialize the agent.
- 2 For each episode:
 - a Reset the environment.
 - b Get the initial observation s_0 from the environment.
 - c Compute the initial action $a_0 = \mu(s_0)$, where $\mu(s)$ is the current policy.
 - d Set the current action to the initial action ($a \leftarrow a_0$), and set the current observation to the initial observation ($s \leftarrow s_0$).
 - e While the episode is not finished or terminated, perform the following steps.
 - i Apply action a to the environment and obtain the next observation s' and the reward r .
 - ii Learn from the experience set (s, a, r, s') .
 - iii Compute the next action $a' = \mu(s')$.
 - iv Update the current action with the next action ($a \leftarrow a'$) and update the current observation with the next observation ($s \leftarrow s'$).
 - v Terminate the episode if the termination conditions defined in the environment are met.
- 3 If the training termination condition is met, terminate training. Otherwise, begin the next episode.

The specifics of how the software performs these steps depend on the configuration of the agent and environment. For instance, resetting the environment at the start of each episode can include randomizing initial state values, if you configure your environment to do so. For more information on agents and their training algorithms, see “Reinforcement Learning Agents” on page 4-2.

Episode Manager

By default, calling the `train` function opens the Reinforcement Learning Episode Manager, which lets you visualize the training progress. The Episode Manager plot shows the reward for each episode (**EpisodeReward**) and a running average reward value (**AverageReward**). Also, for agents that have critics, the plot shows the critic's estimate of the discounted long-term reward at the start of each episode (**EpisodeQ0**). The Episode Manager also displays various episode and training statistics. You can also use the `train` function to return episode and training information.



For agents with a critic, **Episode Q0** is the estimate of the discounted long-term reward at the start of each episode, given the initial observation of the environment. As training progresses, if the critic is well designed, **Episode Q0** approaches the true discounted long-term reward, as shown in the preceding figure.

To turn off the Reinforcement Learning Episode Manager, set the `Plots` option of `rlTrainingOptions` to "none".

Save Candidate Agents

During training, you can save candidate agents that meet conditions you specify in the `SaveAgentCriteria` and `SaveAgentValue` options of your `rlTrainingOptions` object. For instance, you can save any agent whose episode reward exceeds a certain value, even if the overall condition for terminating training is not yet satisfied. For example, save agents when the episode reward is greater than 100.

```
opt = rlTrainingOptions('SaveAgentCriteria','EpisodeReward','SaveAgentValue',100);
```

`train` stores saved agents in a MAT-file in the folder you specify using the `SaveAgentDirectory` option of `rlTrainingOptions`. Saved agents can be useful, for instance, to test candidate agents generated during a long-running training process. For details about saving criteria and saving location, see `rlTrainingOptions`.

After training is complete, you can save the final trained agent from the MATLAB workspace using the `save` function. For example, save the agent `myAgent` to the file `finalAgent.mat` in the current working directory.

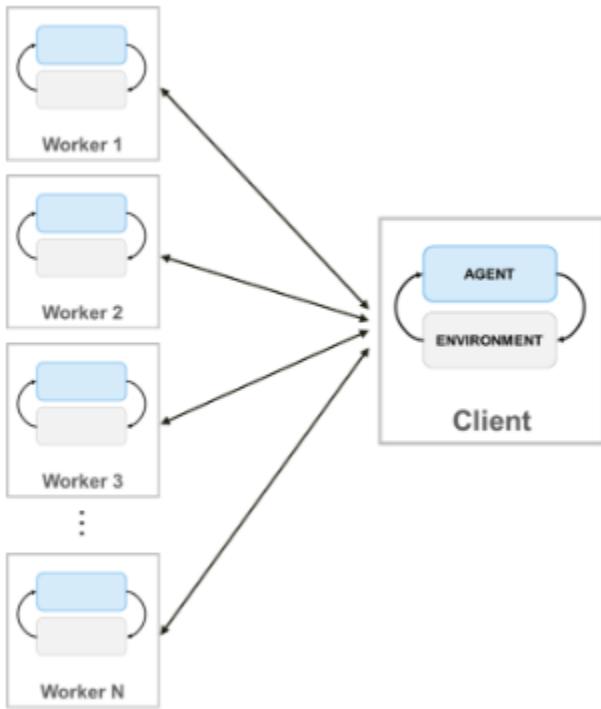
```
save(opt.SaveAgentDirectory + "/finalAgent.mat",'agent')
```

By default, when DDPG and DQN agents are saved, the experience buffer data is not saved. If you plan to further train your saved agent, you can start training with the previous experience buffer as a starting point. In this case, set the `SaveExperienceBufferWithAgent` option to `true`. For some agents, such as those with large experience buffers and image-based observations, the memory required for saving the experience buffer is large. In these cases, you must ensure that enough memory is available for the saved agents.

Parallel Computing

You can accelerate agent training by running parallel training simulations. If you have Parallel Computing Toolbox software, you can run parallel simulations on multicore computers. If you have MATLAB Parallel Server software, you can run parallel simulations on computer clusters or cloud resources.

When you train agents using parallel computing, the host client sends copies of the agent and environment to each parallel worker. Each worker simulates the agent within the environment and sends their simulation data back to the host. The host agent learns from the data sent by the workers and sends the updated policy parameters back to the workers.



To create a parallel pool of N workers, use the following syntax.

```
pool = parpool(N);
```

If you do not create a parallel pool using `parpool`, the `train` function automatically creates one using your default parallel pool preferences. For more information on specifying these preferences, see "Specify Your Parallel Preferences" (Parallel Computing Toolbox).

For off-policy agents, such as DDPG and DQN agents, do not use all of your cores for parallel training. For example, if your CPU has six cores, train with four workers. Doing so provides more resources for the host client to compute gradients based on the experiences sent back from the workers. Limiting the number of workers is not necessary for on-policy agents, such as PG and AC agents, when the gradients are computed on the workers.

For more information on configuring your training to use parallel computing, see the `UseParallel` and `ParallelizationOptions` options in `rlTrainingOptions`.

To benefit from parallel computing, the computational cost for simulating the environment must be relatively expensive compared to the optimization of parameters when sending experiences back to the host. If the simulation of the environment is not expensive enough, the workers idle while waiting for the host to learn and send back updated parameters.

When sending experiences back from the workers, you can improve sample efficiency when the ratio (R) of the environment step complexity to the learning complexity is large. If the environment is fast to simulate (R is small), you are unlikely to get any benefit from experience-based parallelization. If the environment is expensive to simulate but it is also expensive to learn (for example, if the mini-batch size is large), then you are also unlikely to improve sample efficiency. However, in this case, for off-policy agents, you can reduce the mini-batch size to make R larger, which improves sample efficiency.

For an example that trains an agent using parallel computing in MATLAB, see “Train AC Agent to Balance Cart-Pole System Using Parallel Computing” on page 5-84. For an example that trains an agent using parallel computing in Simulink, see “Train DQN Agent for Lane Keeping Assist Using Parallel Computing” on page 5-126.

GPU Acceleration

When using deep neural network function approximators for your actor or critic representations, you can speed up training by performing representation operations on a GPU rather than a CPU. To do so, set the `UseDevice` option to "gpu".

```
opt = rlRepresentationOptions('UseDevice','gpu');
```

The size of any performance improvement depends on your specific application and network configuration.

Validate Trained Policy

To validate your trained agent, you can simulate the agent within the training environment using the `sim` function. To configure the simulation, use `rlSimulationOptions`.

When validating your agent, consider checking how your agent handles the following:

- Changes to simulation initial conditions — To change the model initial conditions, modify the reset function for the environment. For example reset functions, see “Create MATLAB Environment Using Custom Functions” on page 2-33, “Create Custom MATLAB Environment from Template” on page 2-40, and “Create Simulink Environments for Reinforcement Learning” on page 2-5.
- Mismatches between the training and simulation environment dynamics — To check such mismatches, create test environments in the same way that you created the training environment, modifying the environment behavior.

As with parallel training, if you have Parallel Computing Toolbox software, you can run multiple parallel simulations on multicore computers. If you have MATLAB Parallel Server software, you can run multiple parallel simulations on computer clusters or cloud resources. For more information on configuring your simulation to use parallel computing, see `UseParallel` and `ParallelizationOptions` in `rlSimulationOptions`.

Environment Visualization

If your training environment implements the `plot` method, you can visualize the environment behavior during training and simulation. If you call `plot(env)` before training or simulation, where `env` is your environment object, then the visualization updates during training to allow you to visualize the progress of each episode or simulation.

Environment visualization is not supported when training or simulating your agent using parallel computing.

For custom environments, you must implement your own `plot` method. For more information on creating a custom environments with a `plot` function, see “Create Custom MATLAB Environment from Template” on page 2-40.

See Also

`train`

More About

- “Reinforcement Learning Agents” on page 4-2

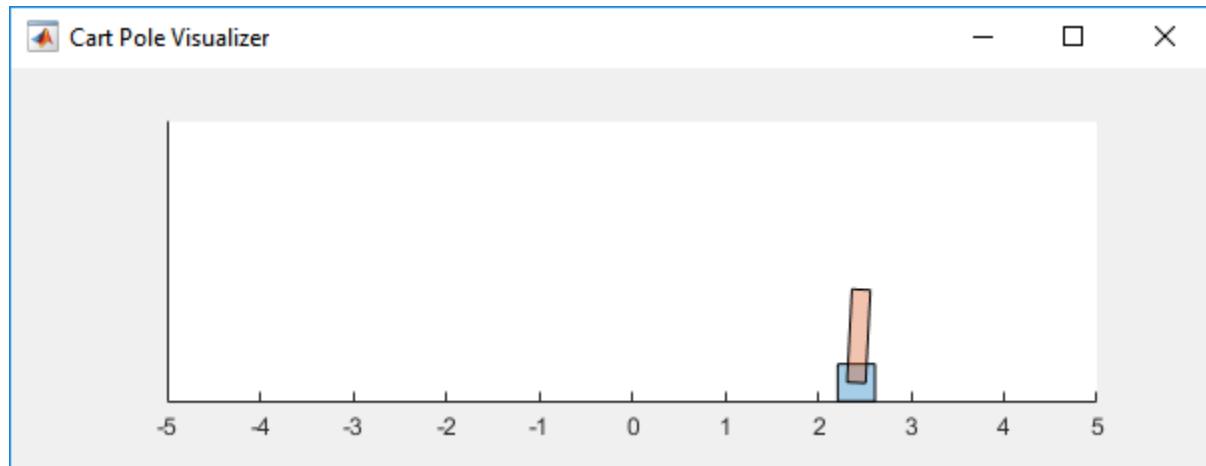
Train DQN Agent to Balance Cart-Pole System

This example shows how to train a deep Q-learning network (DQN) agent to balance a cart-pole system modeled in MATLAB®.

For more information on DQN agents, see “Deep Q-Network Agents” on page 4-8. For an example that trains a DQN agent in Simulink®, see “Train DQN Agent to Swing Up and Balance Pendulum” on page 5-37.

Cart-pole MATLAB Environment

The reinforcement learning environment for this example is a pole attached to an unactuated joint on a cart, which moves along a frictionless track. The training goal is to make the pendulum stand upright without falling over.



For this environment:

- The upward balanced pendulum position is θ radians, and the downward hanging position is π radians
- The pendulum starts upright with an initial angle of ± 0.05 radians
- The force action signal from the agent to the environment is from -10 to 10 N
- The observations from the environment are the position and velocity of the cart, the pendulum angle, and its derivative
- The episode terminates if the pole is more than 12 degrees from vertical, or the cart moves more than 2.4 m from the original position
- A reward of +1 is provided for every time step that the pole remains upright. A penalty of -5 is applied when the pendulum falls.

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv("CartPole-Discrete")
```

```

env =
    CartPoleDiscreteAction with properties:

        Gravity: 9.8000
        MassCart: 1
        MassPole: 0.1000
        Length: 0.5000
        MaxForce: 10
        Ts: 0.0200
    ThetaThresholdRadians: 0.2094
        XThreshold: 2.4000
    RewardForNotFalling: 1
    PenaltyForFalling: -5
        State: [4x1 double]

```

The interface has a discrete action space where the agent can apply one of two possible force values to the cart, -10 or 10 N.

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DQN agent

A DQN agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

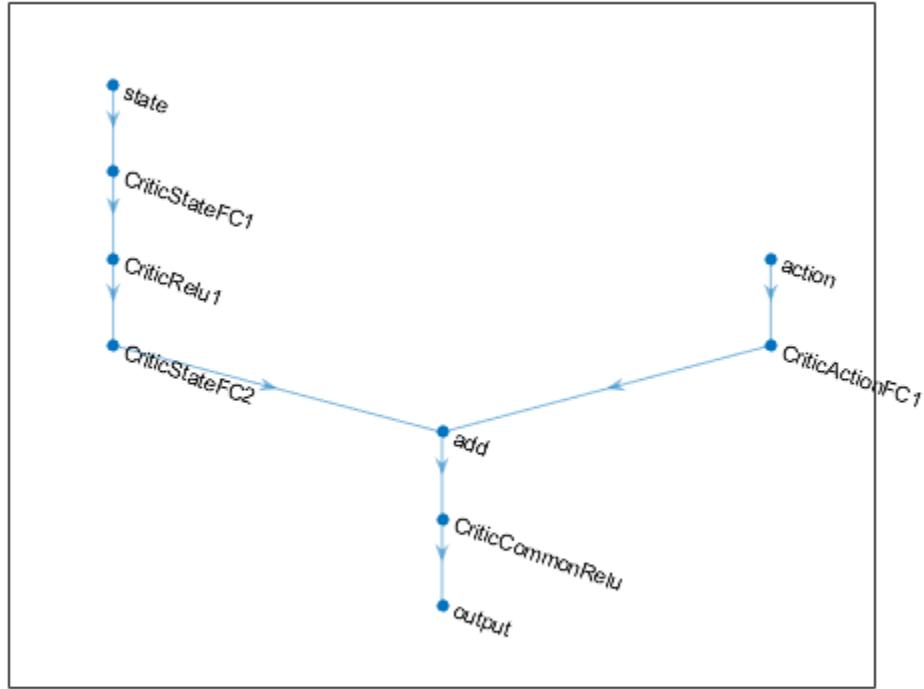
```

statePath = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(24, 'Name', 'CriticStateFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(24, 'Name', 'CriticStateFC2'));
actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(24, 'Name', 'CriticActionFC1'));
commonPath = [
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'CriticCommonRelu')
    fullyConnectedLayer(1, 'Name', 'output'));
criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = addLayers(criticNetwork, commonPath);
criticNetwork = connectLayers(criticNetwork, 'CriticStateFC2', 'add/in1');
criticNetwork = connectLayers(criticNetwork, 'CriticActionFC1', 'add/in2');

```

View the critic network configuration.

```
figure
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate',0.01,'GradientThreshold',1);
```

Create the critic representation using the specified neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
obsInfo = getObservationInfo(env);
actInfo = getActionInfo(env);
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,'Observation',{'state'},'Action',{});
```

To create the DQN agent, first specify the DQN agent options using `rldQNAgentOptions`.

```
agentOpts = rldQNAgentOptions(
    'UseDoubleDQN',false,
    'TargetUpdateMethod','periodic',
    'TargetUpdateFrequency',4,
    'ExperienceBufferLength',100000,
    'DiscountFactor',0.99,
    'MiniBatchSize',256);
```

Then, create the DQN agent using the specified critic representation and agent options. For more information, see `rldQNAgent`.

```
agent = rldQNAgent(critic,agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

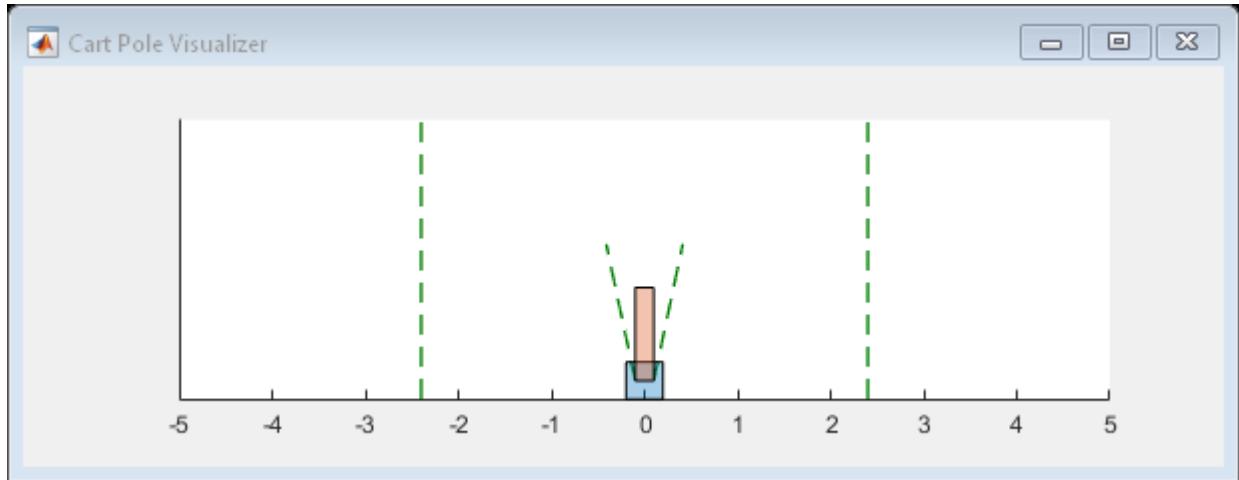
- Run one training session containing at most 1000 episodes, with each episode lasting at most 500 time steps.
- Display the training progress in the Episode Manager dialog box (set the **Plots** option) and disable the command line display (set the **Verbose** option to **false**).
- Stop training when the agent receives an moving average cumulative reward greater than 480. At this point, the agent can balance the cart-pole system in the upright position.

For more information, see [rlTrainingOptions](#).

```
trainOpts = rlTrainingOptions(...
    'MaxEpisodes', 1000, ...
    'MaxStepsPerEpisode', 500, ...
    'Verbose', false, ...
    'Plots', 'training-progress',...
    'StopTrainingCriteria', 'AverageReward',...
    'StopTrainingValue', 480);
```

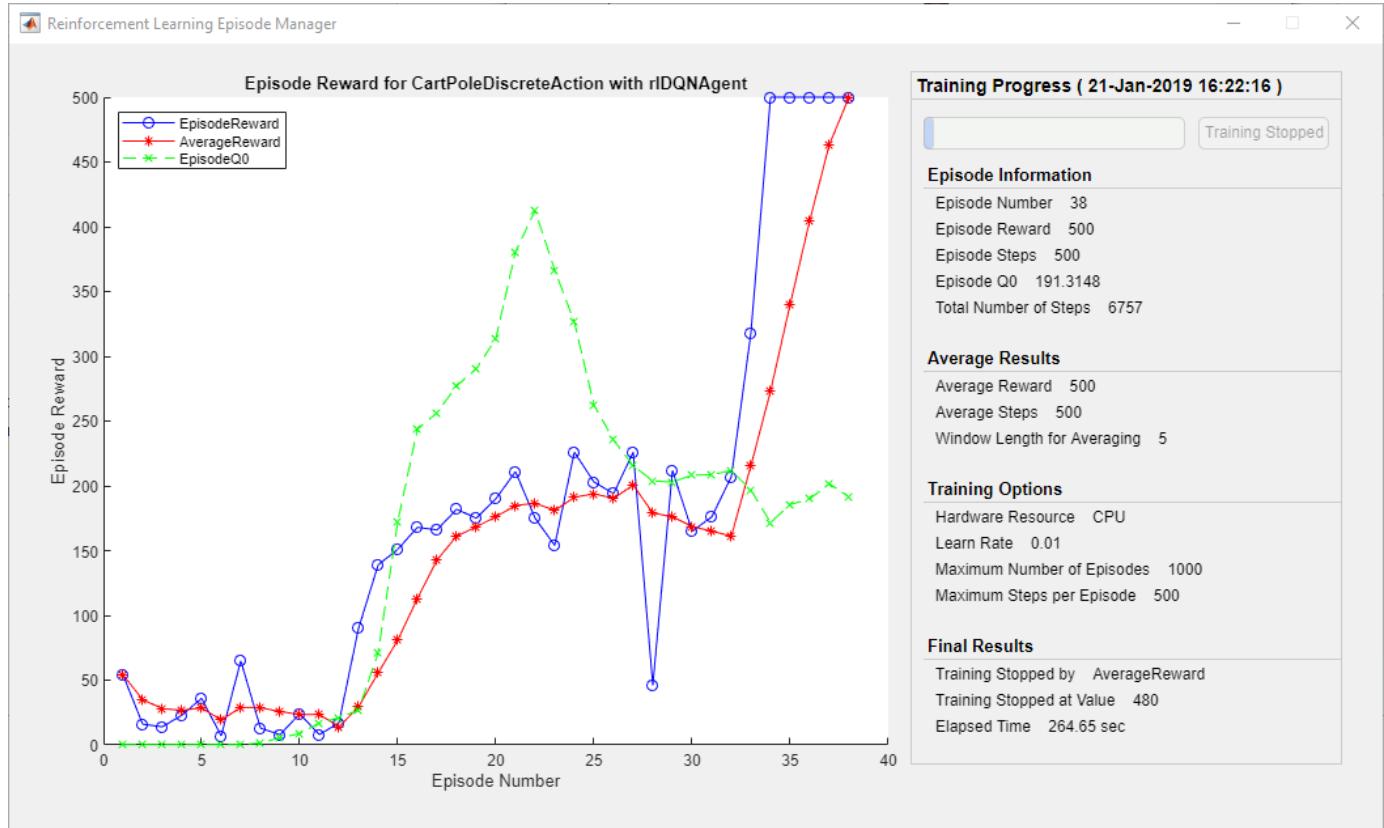
The cart-pole system can be visualized with using the **plot** function during training or simulation.

```
plot(env)
```



Train the agent using the **train** function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting **doTraining** to **false**. To train the agent yourself, set **doTraining** to **true**.

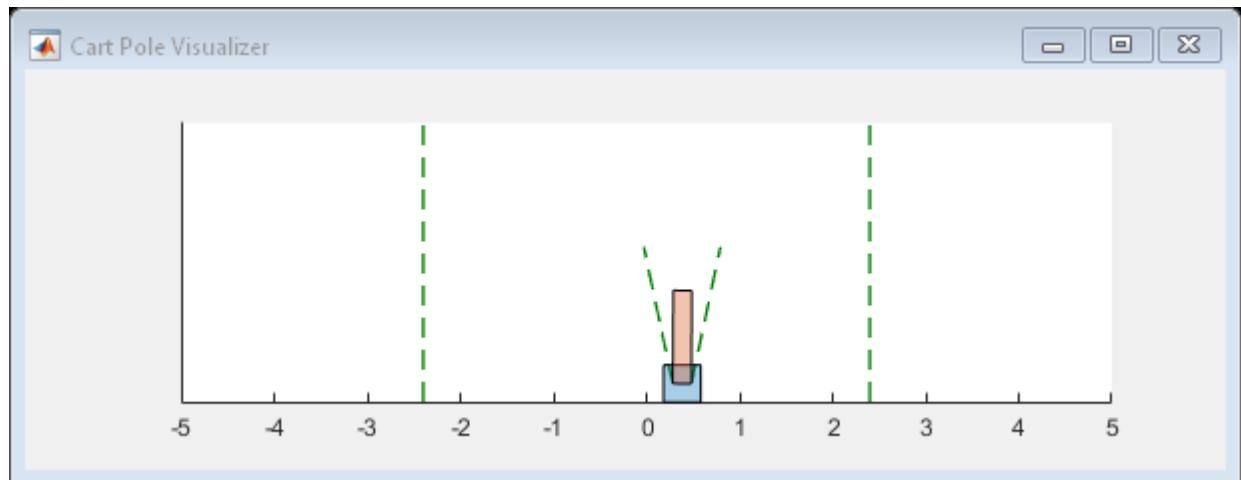
```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('MATLABCartpoleDQN.mat','agent');
end
```



Simulate DQN Agent

To validate the performance of the trained agent, simulate it within the cart-pole environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`. The agent can balance the cart-pole even when simulation time increases to 500.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)
```

```
totalReward = 500
```

See Also

[train](#)

More About

- “Deep Q-Network Agents” on page 4-8
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

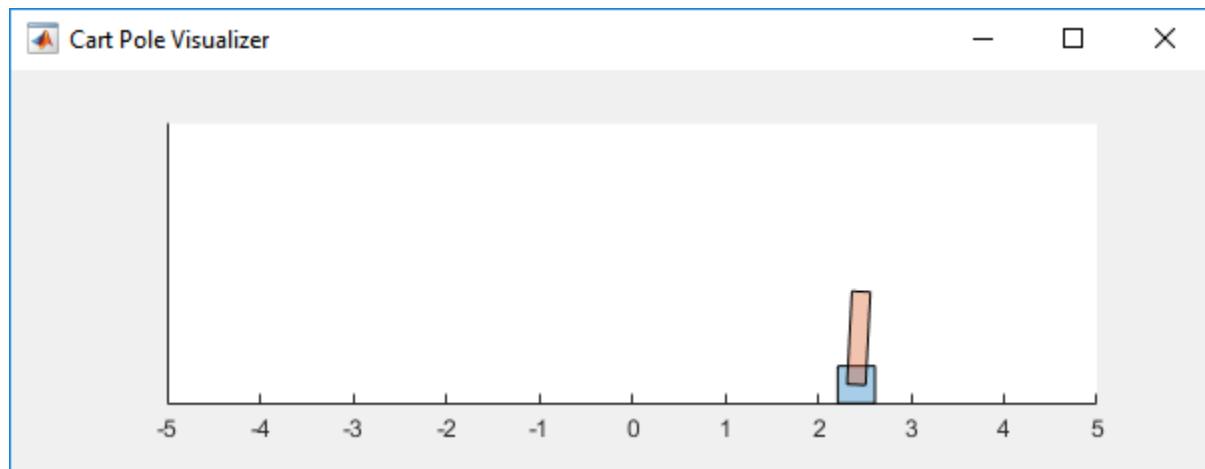
Train PG Agent to Balance Cart-Pole System

This example shows how to train a policy gradient (PG) agent to balance a cart-pole system modeled in MATLAB®. For more information on PG agents, see “Policy Gradient Agents” on page 4-11.

For an example on training a PG agent with baseline, see “Train PG Agent with Baseline to Control Double Integrator System” on page 5-25 example.

Cart-pole MATLAB Environment

The reinforcement learning environment for this example is a pole attached to an unactuated joint on a cart, which moves along a frictionless track. The training goal is to make the pendulum stand upright without falling over.



For this environment:

- The upward balanced pendulum position is 0 radians, and the downward hanging position is π radians
- The pendulum starts upright with an initial angle of +/- 0.05 radians
- The force action signal from the agent to the environment is from -10 to 10 N
- The observations from the environment are the position and velocity of the cart, the pendulum angle, and its derivative
- The episode terminates if the pole is more than 12 degrees from vertical, or the cart moves more than 2.4 m from the original position
- A reward of +1 is provided for every time step that the pole remains upright. A penalty of -5 is applied when the pendulum falls.

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv("CartPole-Discrete")
env =
  CartPoleDiscreteAction with properties:
```

```

    Gravity: 9.8000
    MassCart: 1
    MassPole: 0.1000
    Length: 0.5000
    MaxForce: 10
    Ts: 0.0200
    ThetaThresholdRadians: 0.2094
    XThreshold: 2.4000
    RewardForNotFalling: 1
    PenaltyForFalling: -5
    State: [4x1 double]

```

The interface has a discrete action space where the agent can apply one of two possible force values to the cart, -10 or 10 N.

Obtain the observation and action information from the environment interface.

```

obsInfo = getObservationInfo(env);
numObservations = obsInfo.Dimension(1);
actInfo = getActionInfo(env);

```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create PG agent

A PG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input (the observation) and one output (the action). The output size actor network is 2 since the agent can apply 2 possible actions, 10 or -10. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```

actorNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(2, 'Name', 'fc')
    softmaxLayer('Name', 'actionProb')
];

```

Specify options for the actor representation using `rlRepresentationOptions`.

```
actorOpts = rlRepresentationOptions('LearnRate', 1e-2, 'GradientThreshold', 1);
```

Create the actor representation using the specified deep neural network and options. You must also specify the action and observation information for the critic, which you already obtained from the environment interface. For more information, see `rlStochasticActorRepresentation`.

```
actor = rlStochasticActorRepresentation(actorNetwork, obsInfo, actInfo, 'Observation', {'state'}, act
```

Then, create the agent using the specified actor representation, and the default agent options. For more information, see `rlPGAgent`.

```
agent = rlPGAgent(actor);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

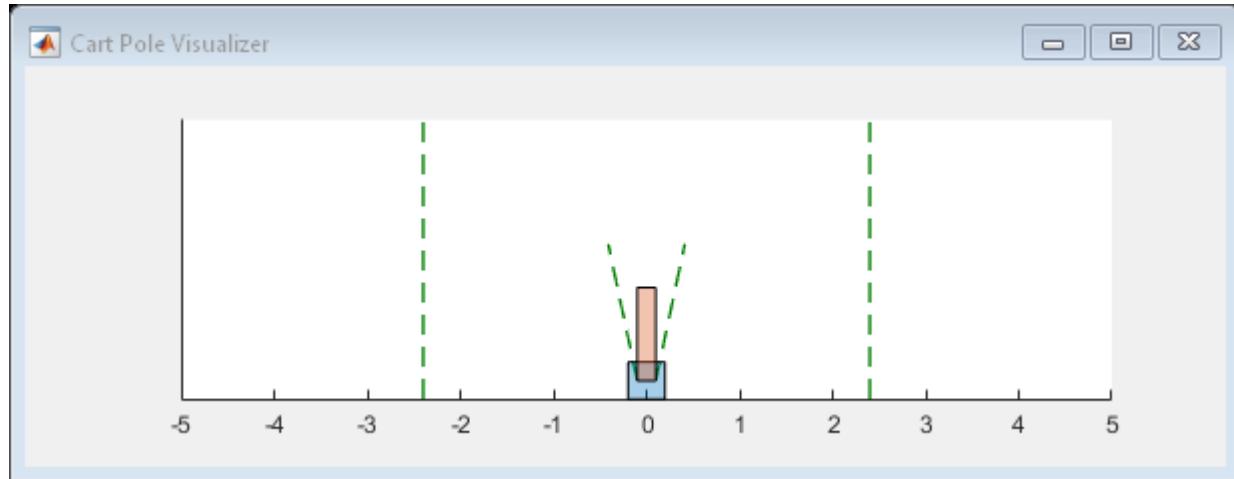
- Run each training episode for at most 1000 episodes, with each episode lasting at most 200 time steps.
- Display the training progress in the Episode Manager dialog box (set the **Plots** option) and disable the command line display (set the **Verbose** option to **false**).
- Stop training when the agent receives an average cumulative reward greater than 195 over 100 consecutive episodes. At this point, the agent can balance the pendulum in the upright position.

For more information, see [rlTrainingOptions](#).

```
trainOpts = rlTrainingOptions(...
    'MaxEpisodes', 1000, ...
    'MaxStepsPerEpisode', 200, ...
    'Verbose', false, ...
    'Plots', 'training-progress',...
    'StopTrainingCriteria', 'AverageReward',...
    'StopTrainingValue', 195, ...
    'ScoreAveragingWindowLength', 100);
```

The cart-pole system can be visualized with using the **plot** function during training or simulation.

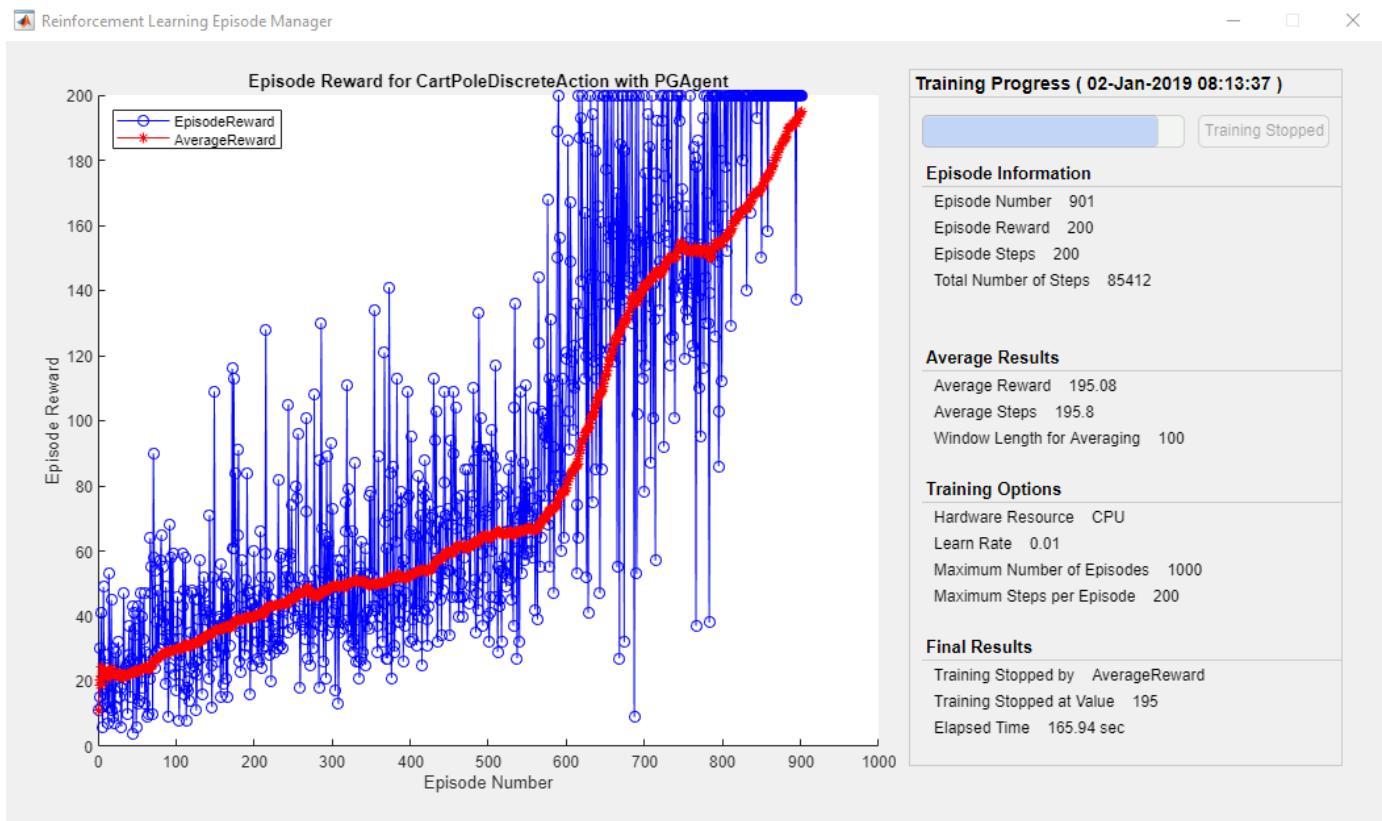
```
plot(env)
```



Train the agent using the **train** function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting **doTraining** to **false**. To train the agent yourself, set **doTraining** to **true**.

```
doTraining = false;

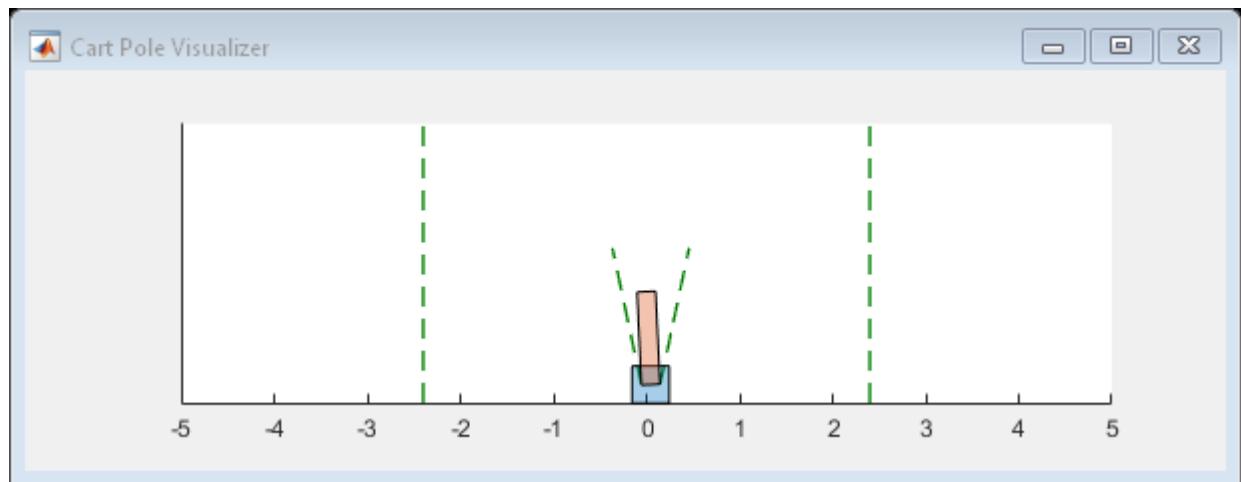
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('MATLABCartpolePG.mat','agent');
end
```



Simulate PG Agent

To validate the performance of the trained agent, simulate it within the cart-pole environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`. The agent can balance the cart-pole even when simulation time increases to 500.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)
```

```
totalReward = 500
```

See Also

[train](#)

More About

- “Policy Gradient Agents” on page 4-11
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

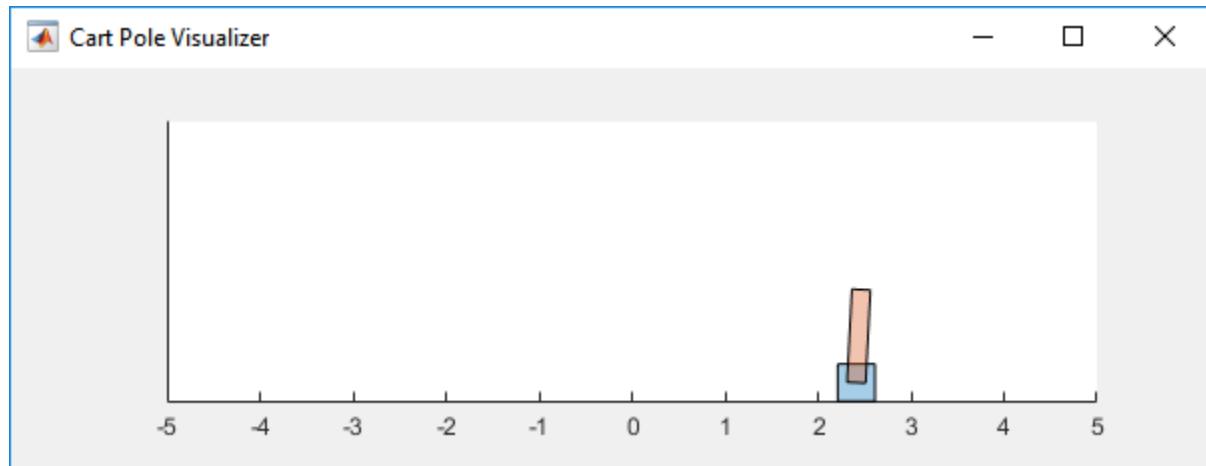
Train AC Agent to Balance Cart-Pole System

This example shows how to train an actor-critic (AC) agent to balance a cart-pole system modeled in MATLAB®.

For more information on AC agents, see “Actor-Critic Agents” on page 4-21. For an example on training AC agent using parallel computing, see “Train AC Agent to Balance Cart-Pole System Using Parallel Computing” on page 5-84 example.

Cart-Pole MATLAB Environment

The reinforcement learning environment for this example is a pole attached to an unactuated joint on a cart, which moves along a frictionless track. The training goal is to make the pendulum stand upright without falling over.



For this environment:

- The upward balanced pendulum position is θ radians, and the downward hanging position is π radians
- The pendulum starts upright with an initial angle of ± 0.05 radians
- The force action signal from the agent to the environment is from -10 to 10 N
- The observations from the environment are the position and velocity of the cart, the pendulum angle, and its derivative
- The episode terminates if the pole is more than 12 degrees from vertical, or the cart moves more than 2.4 m from the original position
- A reward of +1 is provided for every time step that the pole remains upright. A penalty of -5 is applied when the pendulum falls.

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv("CartPole-Discrete")
```

```
env =
    CartPoleDiscreteAction with properties:

        Gravity: 9.8000
        MassCart: 1
        MassPole: 0.1000
        Length: 0.5000
        MaxForce: 10
        Ts: 0.0200
    ThetaThresholdRadians: 0.2094
        XThreshold: 2.4000
    RewardForNotFalling: 1
    PenaltyForFalling: -5
        State: [4x1 double]

env.PenaltyForFalling = -10;
```

The interface has a discrete action space where the agent can apply one of two possible force values to the cart, -10 or 10 N.

Obtain the observation and action information from the environment interface.

```
obsInfo = getObservationInfo(env);
actInfo = getActionInfo(env);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create AC Agent

An AC agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with one input (the observation) and one output (the state value). The input size of the critic network is [4 1 1] since the environment has 4 observations. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
criticNetwork = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(1, 'Name', 'CriticFC')];
```

Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate', 8e-3, 'GradientThreshold', 1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation information for the critic, which you already obtained from the environment interface. For more information, see `rlValueRepresentation`.

```
critic = rlValueRepresentation(criticNetwork, obsInfo, 'Observation', {'state'}, criticOpts);
```

An AC agent decides which action to take given observations using an actor representation. To create the actor, create a deep neural network with one input (the observation) and one output (the action). The output size of the actor network is 2 since the environment has 2 possible actions, -10 and 10.

Construct the actor in a similar manner to the critic. For more information, see `rlStochasticActorRepresentation`.

```

actorNetwork = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(2, 'Name', 'fc')
    softmaxLayer('Name', 'actionProb')];

actorOpts = rlRepresentationOptions('LearnRate', 8e-3, 'GradientThreshold', 1);

actor = rlStochasticActorRepresentation(actorNetwork, obsInfo, actInfo, ...
    'Observation', {'state'}, actorOpts);

```

To create the AC agent, first specify the AC agent options using `rlACAgentOptions`.

```

agentOpts = rlACAgentOptions(...,
    'NumStepsToLookAhead', 32, ...
    'DiscountFactor', 0.99);

```

Then, create the agent using the specified actor representation and the default agent options. For more information, see `rlACAgent`.

```
agent = rlACAgent(actor, critic, agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training episode for at most 1000 episodes, with each episode lasting at most 500 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than 480 over 10 consecutive episodes. At this point, the agent can balance the pendulum in the upright position.

For more information, see `rlTrainingOptions`.

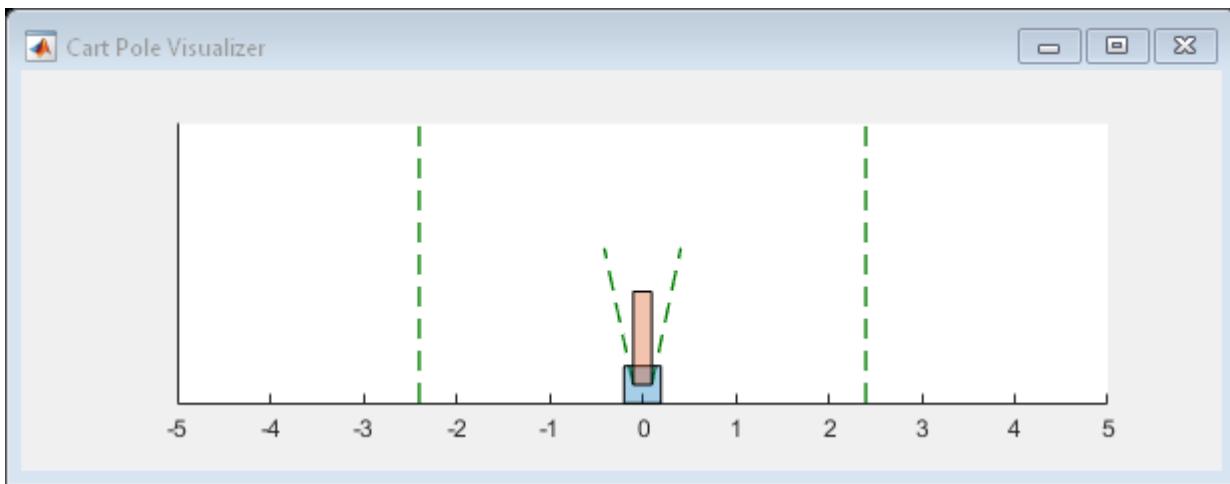
```

trainOpts = rlTrainingOptions(...,
    'MaxEpisodes', 1000, ...
    'MaxStepsPerEpisode', 500, ...
    'Verbose', false, ...
    'Plots', 'training-progress', ...
    'StopTrainingCriteria', 'AverageReward', ...
    'StopTrainingValue', 480, ...
    'ScoreAveragingWindowLength', 10);

```

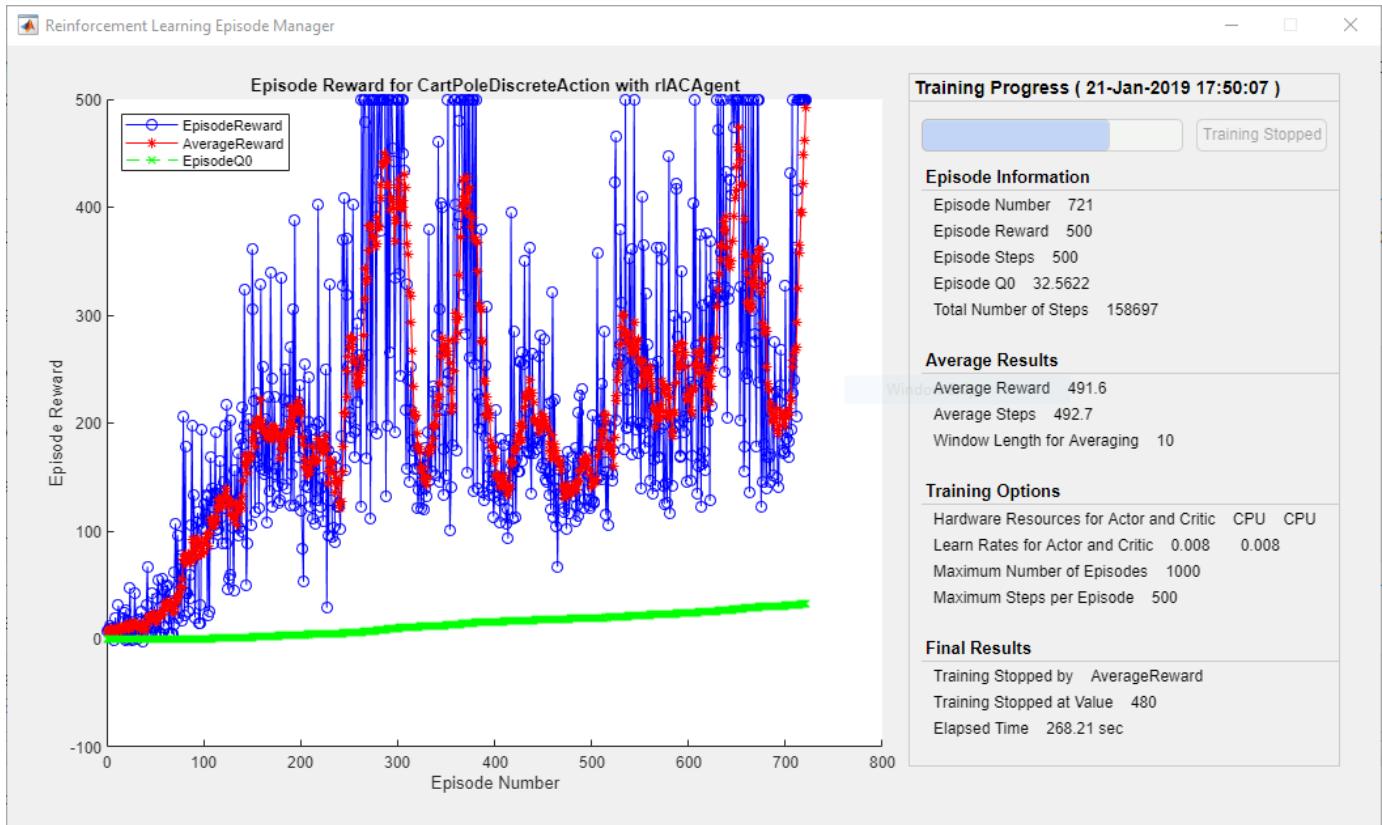
The cart-pole system can be visualized with using the `plot` function during training or simulation.

```
plot(env)
```



Train the agent using the `train` function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

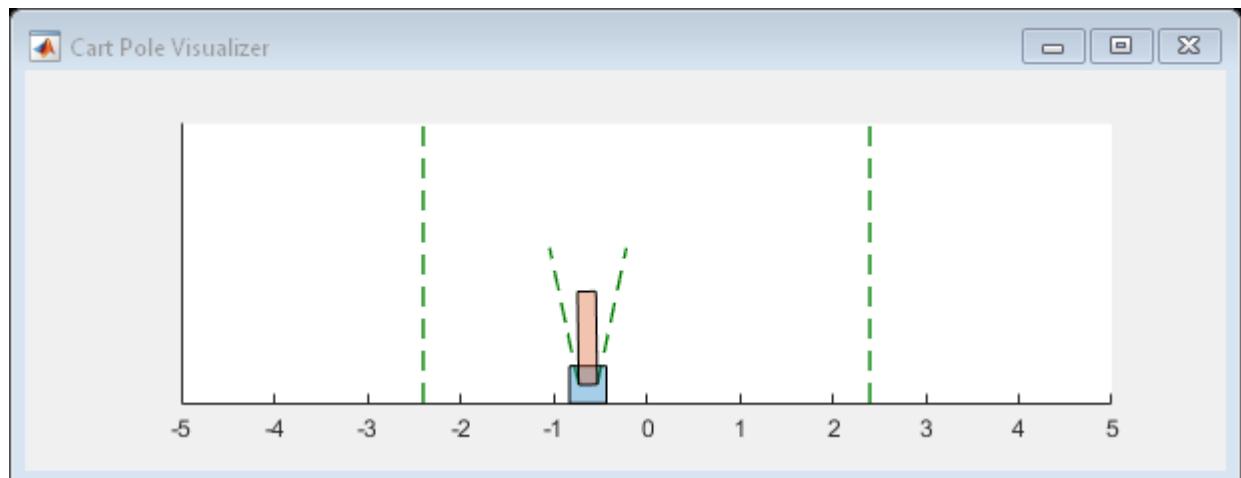
```
doTraining = false;  
  
if doTraining  
    % Train the agent.  
    trainingStats = train(agent,env,trainOpts);  
else  
    % Load pretrained agent for the example.  
    load('MATLABCartpoleAC.mat','agent');  
end
```



Simulate AC Agent

To validate the performance of the trained agent, simulate it within the cart-pole environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)
```

```
totalReward = 500
```

See Also

[train](#)

More About

- “Actor-Critic Agents” on page 4-21
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

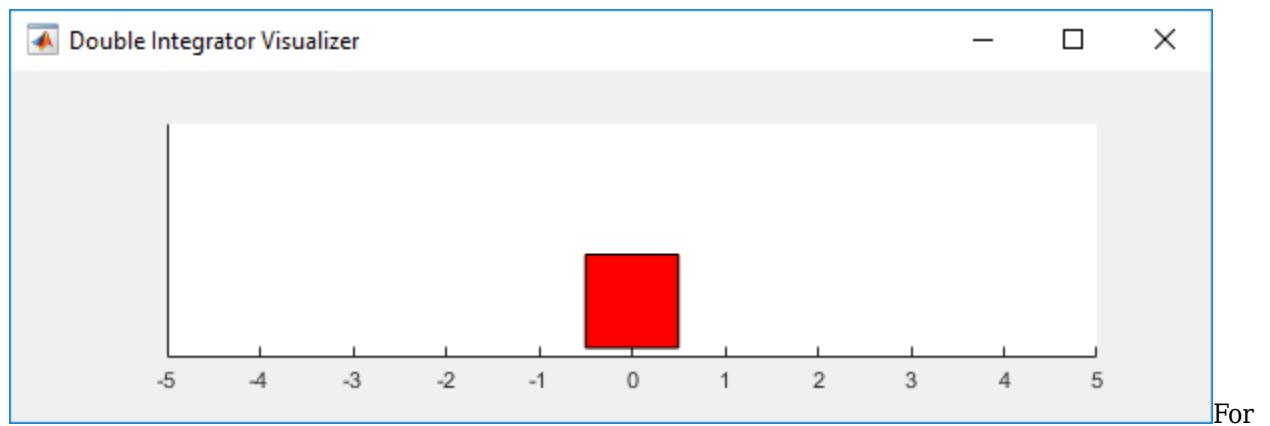
Train PG Agent with Baseline to Control Double Integrator System

This example shows how to train a policy gradient (PG) agent with baseline to control a second-order dynamic system modeled in MATLAB®.

For more information on the basic PG agent with no baseline, see the example “Train PG Agent to Balance Cart-Pole System” on page 5-14.

Double Integrator MATLAB Environment

The reinforcement learning environment for this example is a second-order system is a double integrator with a gain. The training goal is to control the position of a mass in a second-order system by applying a force input.



For this environment:

- The mass starts at initial position of +/- 2 units. The force action signal from the agent to the environment is from -2 to 2 N.
- The observations from the environment are the position and velocity of the mass.
- The episode terminates if the mass moves more than 5 m from the original position or if $|x| < 0.01$
- The reward r_t , provided at every time step, is a discretization of $r(t)$:

$$r(t) = -(x(t)' Q x(t) + u(t)' R u(t))$$

where:

- x is the state vector of the mass.
- u is the force applied to the mass.
- Q is the weights on the control performance. $Q = [10 \ 0; \ 0 \ 1]$
- R is the weight on the control effort. $R = 0.01$

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Double Integrator MATLAB Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv("DoubleIntegrator-Discrete")  
  
env =  
    DoubleIntegratorDiscreteAction with properties:  
  
        Gain: 1  
        Ts: 0.1000  
        MaxDistance: 5  
        GoalThreshold: 0.0100  
        Q: [2x2 double]  
        R: 0.0100  
        MaxForce: 2  
        State: [2x1 double]
```

The interface has a discrete action space where the agent can apply one of three possible force values to the mass: -2, 0 or 2 N.

Obtain the observation and action information from the environment interface.

```
obsInfo = getObservationInfo(env);  
numObservations = obsInfo.Dimension(1);  
actInfo = getActionInfo(env);  
numActions = numel(actInfo.Elements);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create PG agent actor

A PG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input (the observation) and one output (the action). For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
actorNetwork = [  
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'state')  
    fullyConnectedLayer(numActions, 'Name', 'action', 'BiasLearnRateFactor', 0)];
```

Specify options for the actor representation using `rlRepresentationOptions`.

```
actorOpts = rlRepresentationOptions('LearnRate', 5e-3, 'GradientThreshold', 1);
```

Create the actor representation using the specified deep neural network and options. You must also specify the action and observation information for the critic, which you already obtained from the environment interface. For more information, see `rlStochasticActorRepresentation`.

```
actor = rlStochasticActorRepresentation(actorNetwork, obsInfo, actInfo, 'Observation', {'state'}, act
```

Create PG agent baseline

A baseline varies with state can reduce the variance of the expected value of the update and thus reduce the speed of learning for PG agent. A possible choice for the baseline is an estimate of the state value function [1].

In this case, the baseline representation is a deep neural network with one input (the state) and one output (the state value).

Construct the baseline in a similar manner to the actor.

```
baselineNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(8, 'Name', 'BaselineFC')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(1, 'Name', 'BaselineFC2', 'BiasLearnRateFactor', 0)];
baselineOpts = rlRepresentationOptions('LearnRate', 5e-3, 'GradientThreshold', 1);
baseline = rlValueRepresentation(baselineNetwork, obsInfo, 'Observation', {'state'}, baselineOpts);
```

To create the PG agent with baseline, specify the PG agent options with `UseBaseline` option set to true using `rlPGAgentOptions`.

```
agentOpts = rlPGAgentOptions(
    'UseBaseline', true,
    'DiscountFactor', 0.99);
```

Then, create the agent using the specified actor representation, critic representation and agent options. For more information, see `rlPGAgent`.

```
agent = rlPGAgent(actor, baseline, agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

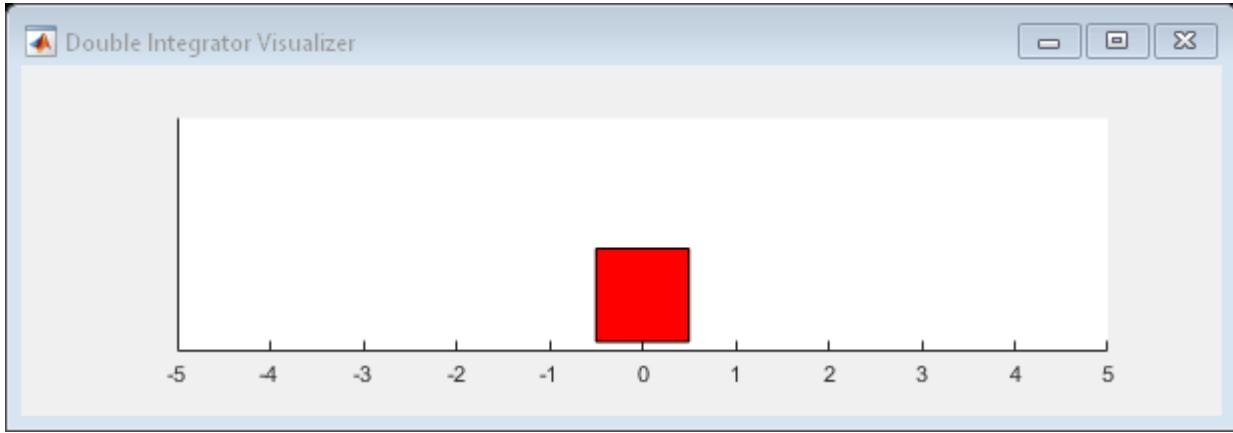
- Run at most 1000 episodes, with each episode lasting at most 200 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option).
- Stop training when the agent receives a moving average cumulative reward greater than -45. At this point, the agent can control the position of the mass using minimal control effort.

For more information, see `rlTrainingOptions`.

```
trainOpts = rlTrainingOptions(
    'MaxEpisodes', 1000,
    'MaxStepsPerEpisode', 200,
    'Verbose', false,
    'Plots', 'training-progress',
    'StopTrainingCriteria', 'AverageReward',
    'StopTrainingValue', -45);
```

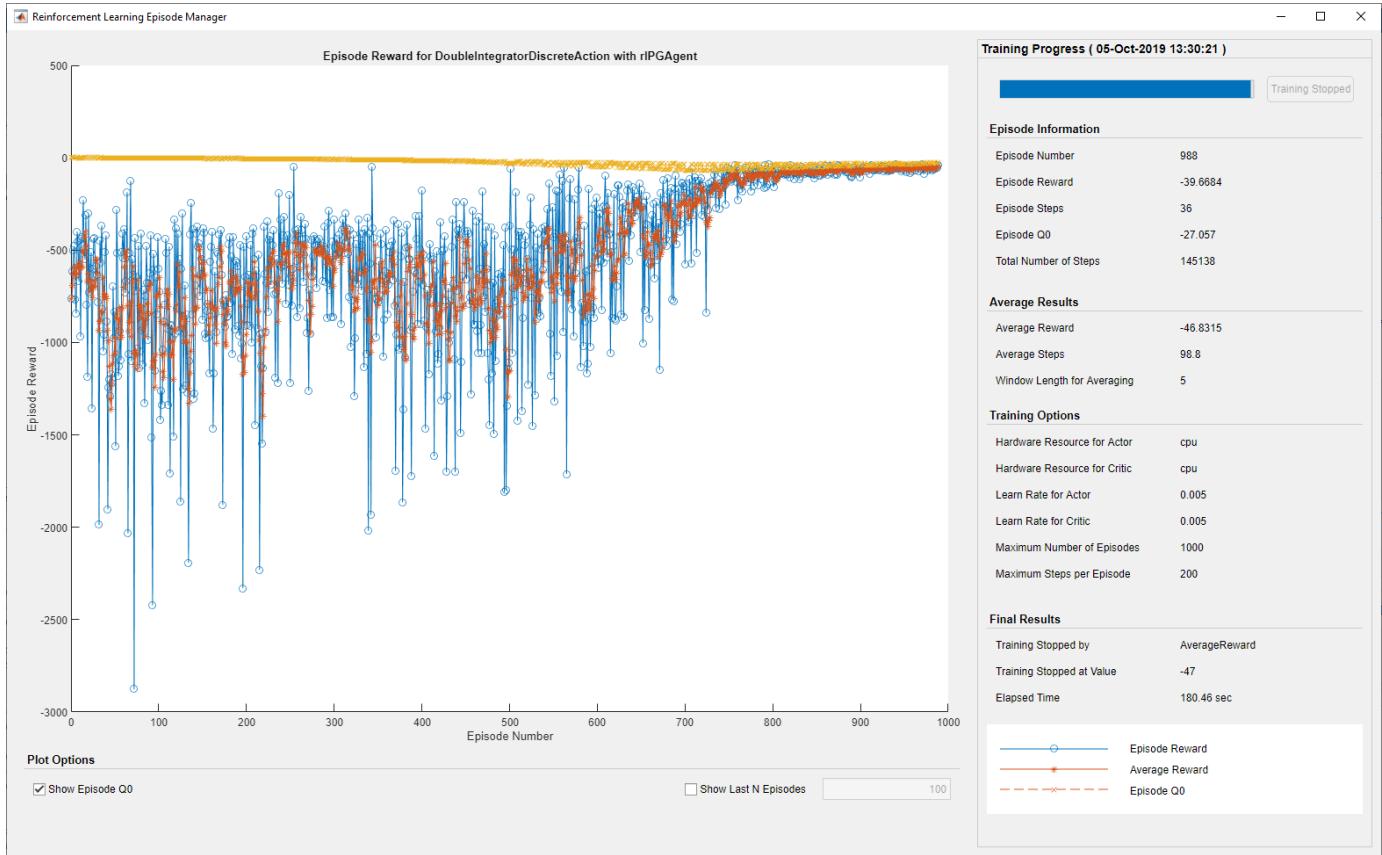
The double integrator system can be visualized with `plot(env)` during training or simulation.

```
plot(env)
```



Train the agent using the `train` function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

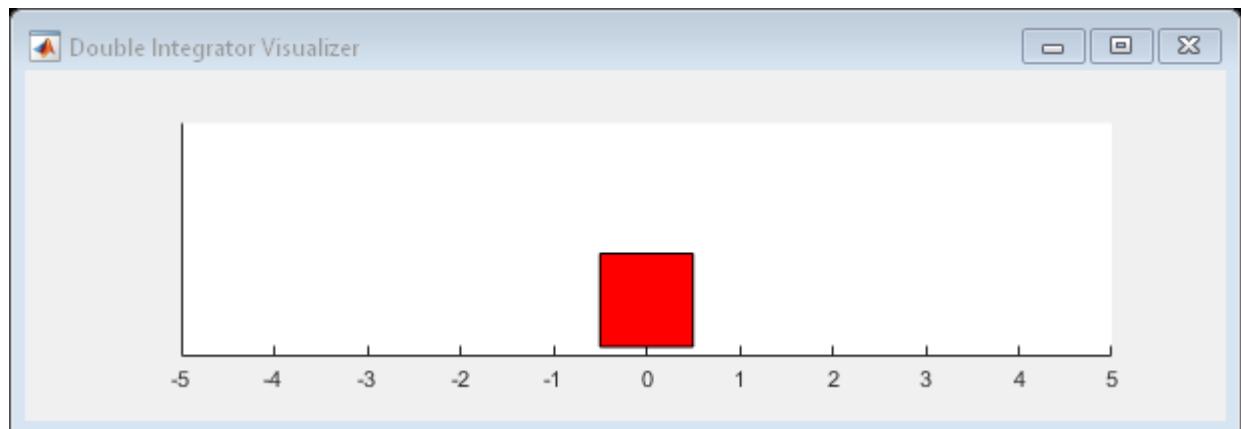
```
doTraining = false;  
  
if doTraining  
    % Train the agent.  
    trainingStats = train(agent,env,trainOpts);  
else  
    % Load pretrained parameters for the example.  
    load('DoubleIntegPGBaselineParams.mat');  
    iLoadPretrainedParams(agent,actorParams,criticParams);  
end
```



Simulate PG Agent

To validate the performance of the trained agent, simulate it within the double integrator environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)
```

```
totalReward = -43.2169
```

Local function

Function to update agent with pretrained parameters

```
function iLoadPretrainedParams(agent,actorParams,criticParams)
    actor = getActor(agent);
    pretrainedActor = setLearnableParameters(actor, actorParams);
    critic = getCritic(agent);
    pretrainedCritic = setLearnableParameters(critic, criticParams);

    setActor(agent,pretrainedActor);
    setCritic(agent,pretrainedCritic);
end
```

References

[1] Sutton, Barto. "Reinforcement Learning: An Introduction," The MIT Press, Cambridge, 2nd Edition, p. 330, 2018.

See Also

[rlPAGAgent](#)

More About

- "Policy Gradient Agents" on page 4-11
- "Train PG Agent to Balance Cart-Pole System" on page 5-14
- "Train Reinforcement Learning Agents" on page 5-2
- "Create Policy and Value Function Representations" on page 3-2

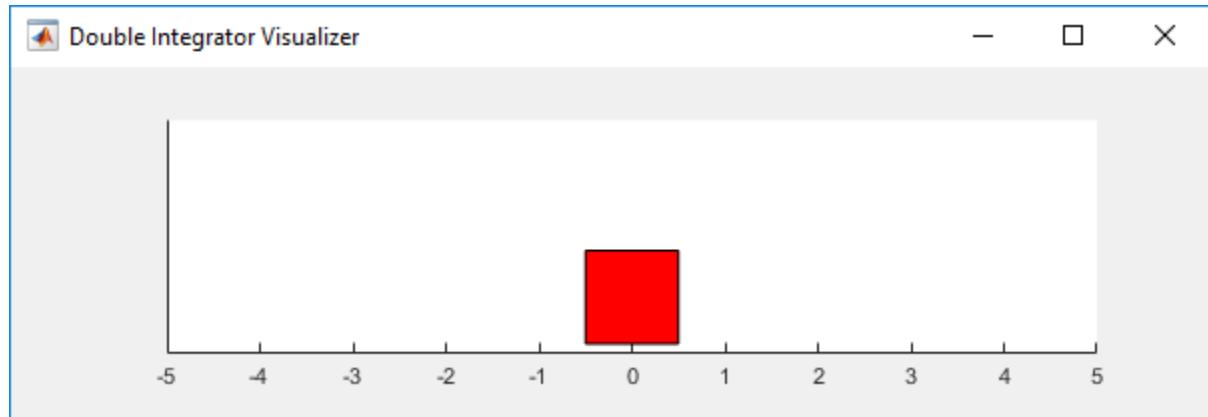
Train DDPG Agent to Control Double Integrator System

This example shows how to train a deep deterministic policy gradient (DDPG) agent to control a second-order dynamic system modeled in MATLAB®.

For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14. For an example that trains a DDPG agent in Simulink®, see “Train DDPG Agent to Swing Up and Balance Pendulum” on page 5-43.

Double Integrator MATLAB Environment

The reinforcement learning environment for this example is a second-order system is a double integrator with a gain. The training goal is to control the position of a mass in a second-order system by applying a force input.



For this environment:

- The mass starts at initial position of +/- 4 units.
- The force action signal from the agent to the environment is from -2 to 2 N.
- The observations from the environment are the position and velocity of the mass.
- The episode terminates if the mass moves more than 5 m from the original position or if $|x| < 0.01$
- The reward r_t , provided at every time step, is a discretization of $r(t)$:

$$r(t) = -(x(t)' Q x(t) + u(t)' R u(t))$$

where:

- x is the state vector of the mass.
- u is the force applied to the mass.
- Q is the weights on the control performance. $Q = [10 \ 0; 0 \ 1]$
- R is the weight on the control effort. $R = 0.01$

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv("DoubleIntegrator-Continuous")  
  
env =  
    DoubleIntegratorContinuousAction with properties:  
  
        Gain: 1  
        Ts: 0.1000  
        MaxDistance: 5  
        GoalThreshold: 0.0100  
        Q: [2x2 double]  
        R: 0.0100  
        MaxForce: Inf  
        State: [2x1 double]
```

```
env.MaxForce = Inf;
```

The interface has a continuous action space where the agent can apply force values from -Inf to Inf to the mass.

Obtain the observation and action information from the environment interface.

```
obsInfo = getObservationInfo(env);  
numObservations = obsInfo.Dimension(1);  
actInfo = getActionInfo(env);  
numActions = numel(actInfo);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

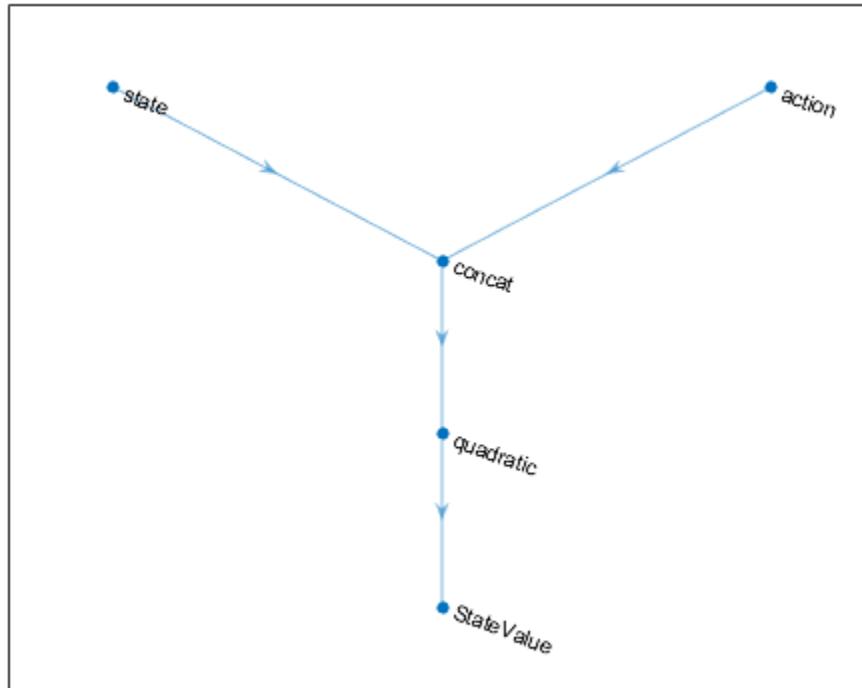
Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
statePath = imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'state');  
actionPath = imageInputLayer([numActions 1 1], 'Normalization', 'none', 'Name', 'action');  
commonPath = [concatenationLayer(1,2, 'Name', 'concat')  
    quadraticLayer('Name', 'quadratic')  
    fullyConnectedLayer(1, 'Name', 'StateValue', 'BiasLearnRateFactor', 0, 'Bias', 0)];  
  
criticNetwork = layerGraph(statePath);  
criticNetwork = addLayers(criticNetwork, actionPath);  
criticNetwork = addLayers(criticNetwork, commonPath);  
  
criticNetwork = connectLayers(criticNetwork, 'state', 'concat/in1');  
criticNetwork = connectLayers(criticNetwork, 'action', 'concat/in2');
```

View the critic network configuration.

```
figure  
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate',5e-3,'GradientThreshold',1);
```

Create the critic representation using the specified neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,'Observation',{'state'},'Action',{...})
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input, the observation, and one output, the action.

Construct the actor similarly to the critic.

```
actorNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization','none','Name','state')
    fullyConnectedLayer(numActions, 'Name','action','BiasLearnRateFactor',0,'Bias',0)];
actorOpts = rlRepresentationOptions('LearnRate',1e-04,'GradientThreshold',1);
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,'Observation',{'state'},'Action',{...})
```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```
agentOpts = rlDDPGAgentOptions(...,'SampleTime',env.Ts,...)
```

```
'TargetSmoothFactor',1e-3,...  
'ExperienceBufferLength',1e6,...  
'DiscountFactor',0.99,...  
'MiniBatchSize',32);  
agent0pts.NoiseOptions.Variance = 0.3;  
agent0pts.NoiseOptions.VarianceDecayRate = 1e-6;
```

Then, create the DDPG agent using the specified actor representation, critic representation and agent options. For more information, see [rlDDPGAgent](#).

```
agent = rlDDPGAgent(actor,critic,agent0pts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

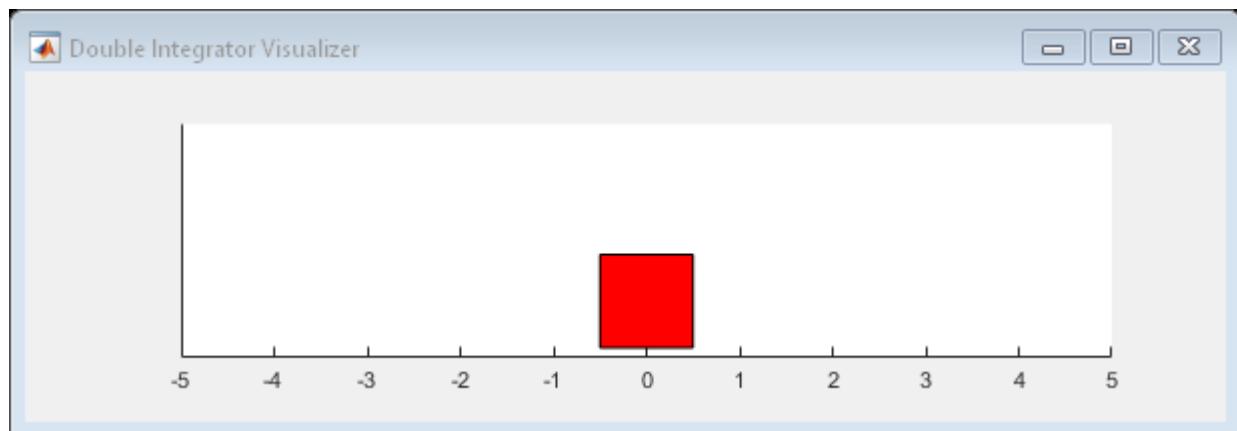
- Run at most 1000 episodes in the training session, with each episode lasting at most 200 time steps.
- Display the training progress in the Episode Manager dialog box (set the **Plots** option) and disable the command line display (set the **Verbose** option).
- Stop training when the agent receives an moving average cumulative reward greater than -66. At this point, the agent can control the position of the mass using minimal control effort.

For more information, see [rlTrainingOptions](#).

```
trainOpts = rlTrainingOptions(...  
    'MaxEpisodes', 5000, ...  
    'MaxStepsPerEpisode', 200, ...  
    'Verbose', false, ...  
    'Plots','training-progress',...  
    'StopTrainingCriteria','AverageReward',...  
    'StopTrainingValue',-66);
```

The double integrator system can be visualized with `plot(env)` during training or simulation.

```
plot(env)
```

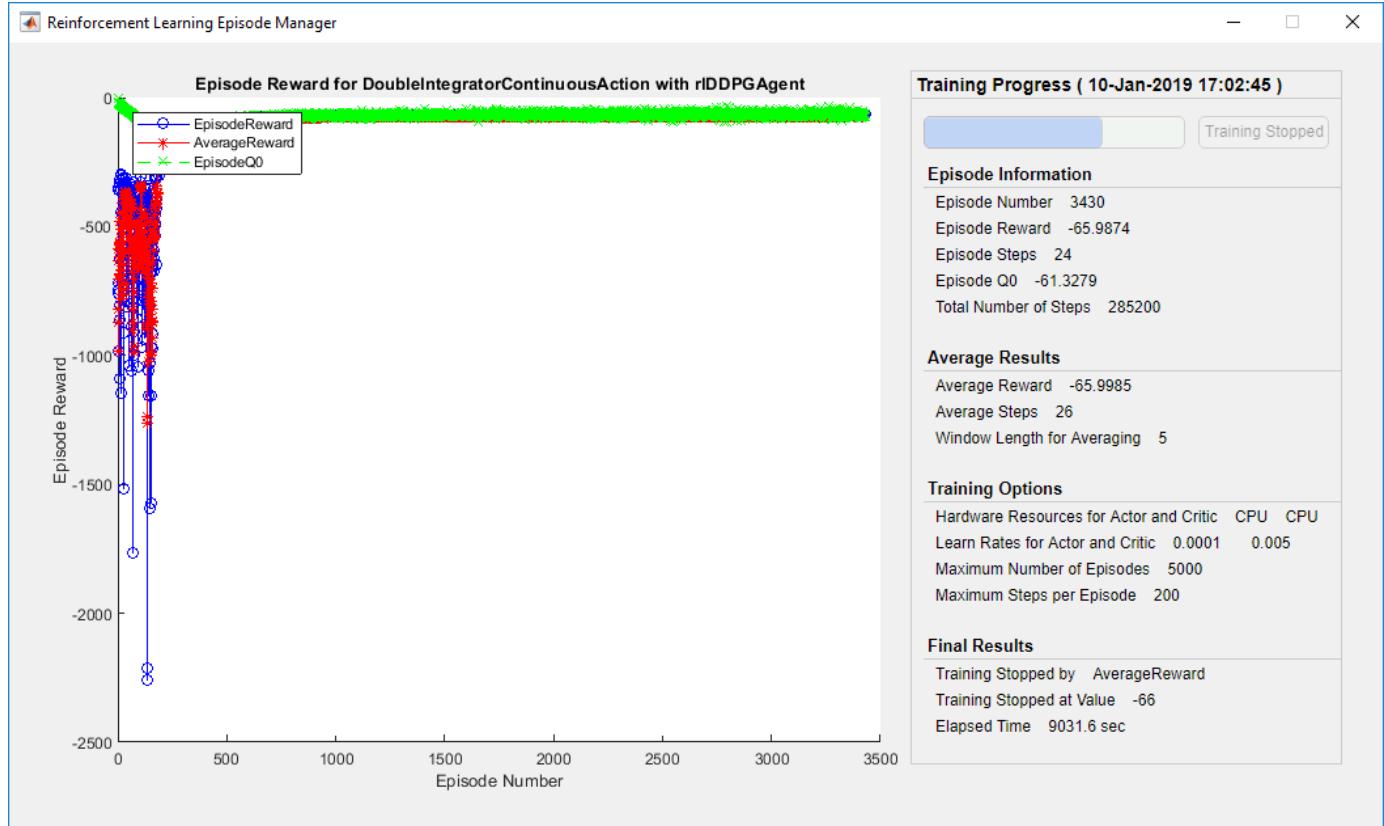


Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```

doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('DoubleIntegDDPG.mat','agent');
end

```



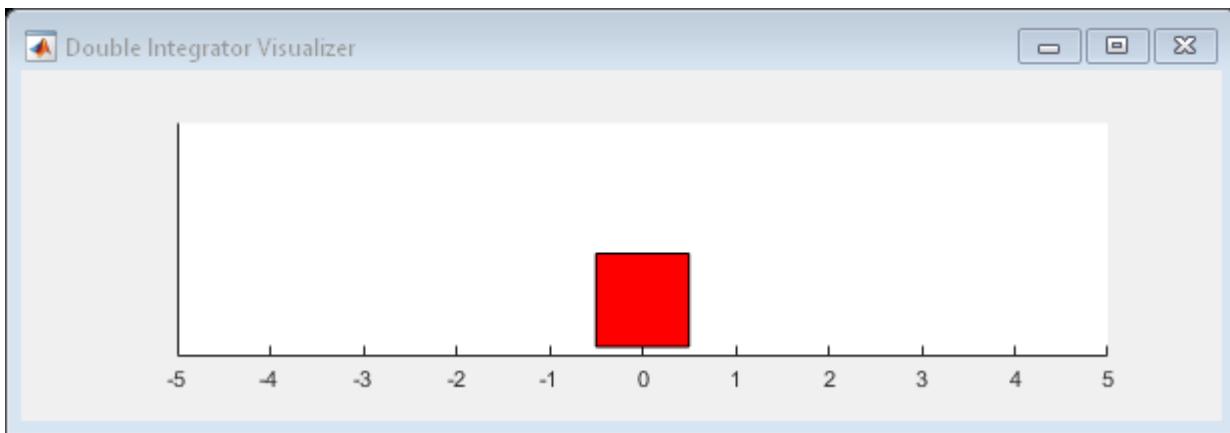
Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the double integrator environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```

simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);

```



```
totalReward = sum(experience.Reward)
```

```
totalReward = single  
-65.9933
```

See Also

[train](#)

More About

- “Deep Deterministic Policy Gradient Agents” on page 4-14
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Train DQN Agent to Swing Up and Balance Pendulum

This example shows how to train a deep Q-learning network (DQN) agent to swing up and balance a pendulum modeled in Simulink®.

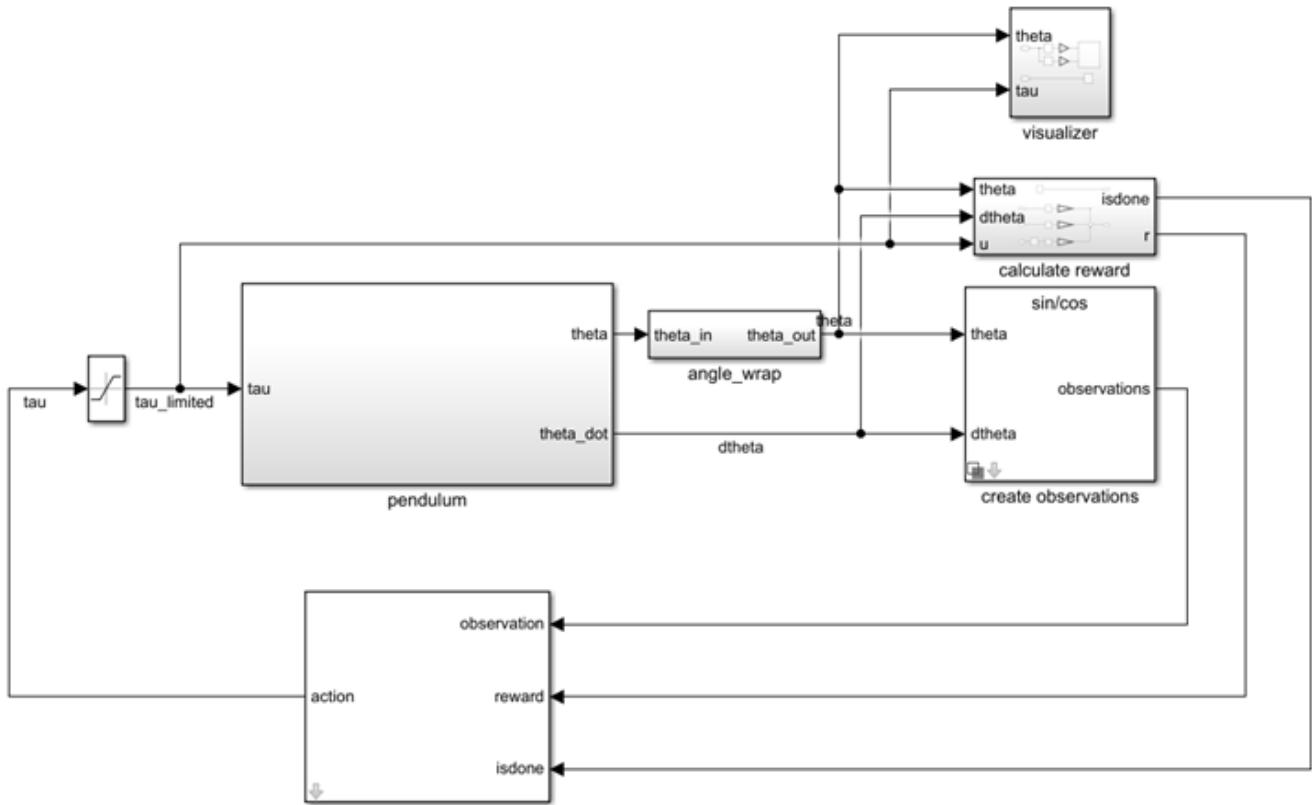
For more information on DQN agents, see “Deep Q-Network Agents” on page 4-8. For an example that trains a DQN agent in MATLAB®, see “Train DQN Agent to Balance Cart-Pole System” on page 5-8.

Pendulum Swing-up Model

The reinforcement learning environment for this example is a simple frictionless pendulum that is initially hanging in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

Open the model.

```
mdl = 'rlSimplePendulumModel';
open_system(mdl)
```



For this model:

- The upward balanced pendulum position is $\theta = 0$ radians, and the downward hanging position is $\theta = \pi$ radians.
- The torque action signal from the agent to the environment is from -2 to 2 Nm.

- The observations from the environment are the sine of the pendulum angle, the cosine of the pendulum angle, and the pendulum angle derivative.
- The reward r_t , provided at every timestep, is:

$$r_t = -(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001u_{t-1}^2)$$

where:

- θ_t is the angle of displacement from the upright position
- $\dot{\theta}_t$ is the derivative of the displacement angle
- u_{t-1} is the control effort from the previous time step.

For more information on this model, see “Load Predefined Simulink Environments” on page 2-22.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv('SimplePendulumModel-Discrete')

env =
SimulinkEnvWithAgent with properties:

    Model: "rlSimplePendulumModel"
    AgentBlock: "rlSimplePendulumModel/RL Agent"
    ResetFcn: []
    UseFastRestart: 'on'
```

The interface has a discrete action space where the agent can apply one of three possible torque values to the pendulum: -2, 0, or 2 Nm.

To define the initial condition of the pendulum as hanging downward, specify an environment reset function using an anonymous function handle. This reset function sets the model workspace variable `theta0` to `pi`.

```
env.ResetFcn = @(in)setVariable(in,'theta0',pi,'Workspace',mdl);
```

Specify the simulation time `Tf` and the agent sample time `Ts` in seconds

```
Ts = 0.05;
Tf = 20;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DQN agent

A DQN agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. The input size of the state path is [3 1 1] since the environment provides 3 observations. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```

statePath = [
    imageInputLayer([3 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(24, 'Name', 'CriticStateFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(48, 'Name', 'CriticStateFC2')];
actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(48, 'Name', 'CriticActionFC1', 'BiasLearnRateFactor', 0)];
commonPath = [
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'CriticCommonRelu')
    fullyConnectedLayer(1, 'Name', 'output')];
criticNetwork = layerGraph();
criticNetwork = addLayers(criticNetwork, statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = addLayers(criticNetwork, commonPath);
criticNetwork = connectLayers(criticNetwork, 'CriticStateFC2', 'add/in1');
criticNetwork = connectLayers(criticNetwork, 'CriticActionFC1', 'add/in2');

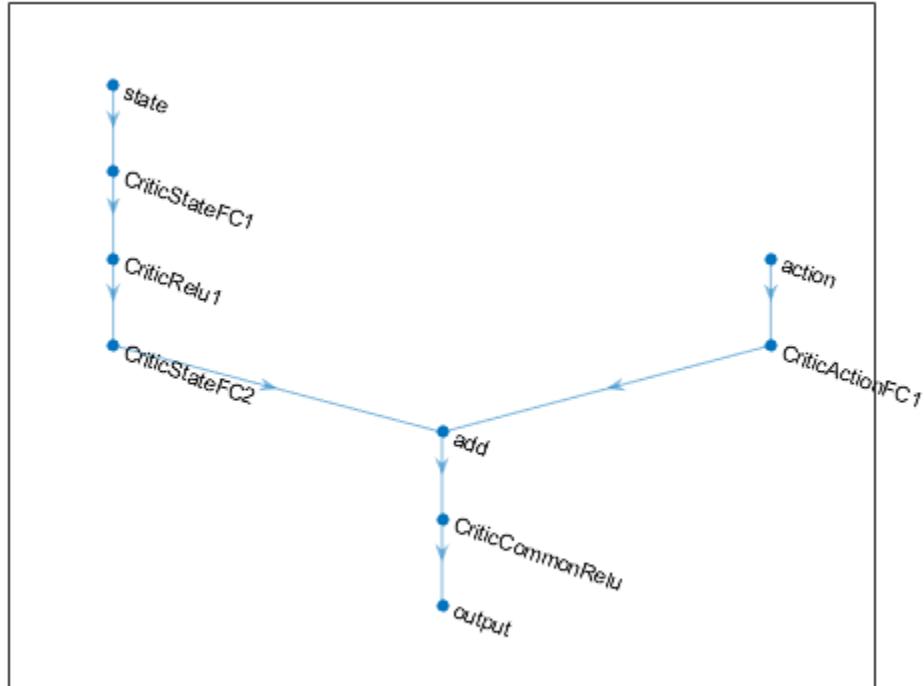
```

View the critic network configuration.

```

figure
plot(criticNetwork)

```



Specify options for the critic representation using `rlRepresentationOptions`.

```

criticOptions = rlRepresentationOptions('LearnRate', 0.01, 'GradientThreshold', 1);

```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
obsInfo = getObservationInfo(env);
actInfo = getActionInfo(env);
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,...  
    'Observation',{ 'state' },'Action',{ 'action' },criticOptions);
```

To create the DQN agent, first specify the DQN agent options using `rlDQNAgentOptions`.

```
agentOptions = rlDQNAgentOptions(...  
    'SampleTime',Ts,...  
    'TargetSmoothFactor',1e-3,...  
    'ExperienceBufferLength',3000,...  
    'UseDoubleDQN',false,...  
    'DiscountFactor',0.9,...  
    'MiniBatchSize',64);
```

Then, create the DQN agent using the specified critic representation and agent options. For more information, see `rlDQNAgent`.

```
agent = rlDQNAgent(critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training for at most `1000` episodes, with each episode lasting at most `500` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than `-1100` over five consecutive episodes. At this point, the agent can quickly balance the pendulum in the upright position using minimal control effort.
- Save a copy of the agent for each episode where the cumulative reward is greater than `-1100`.

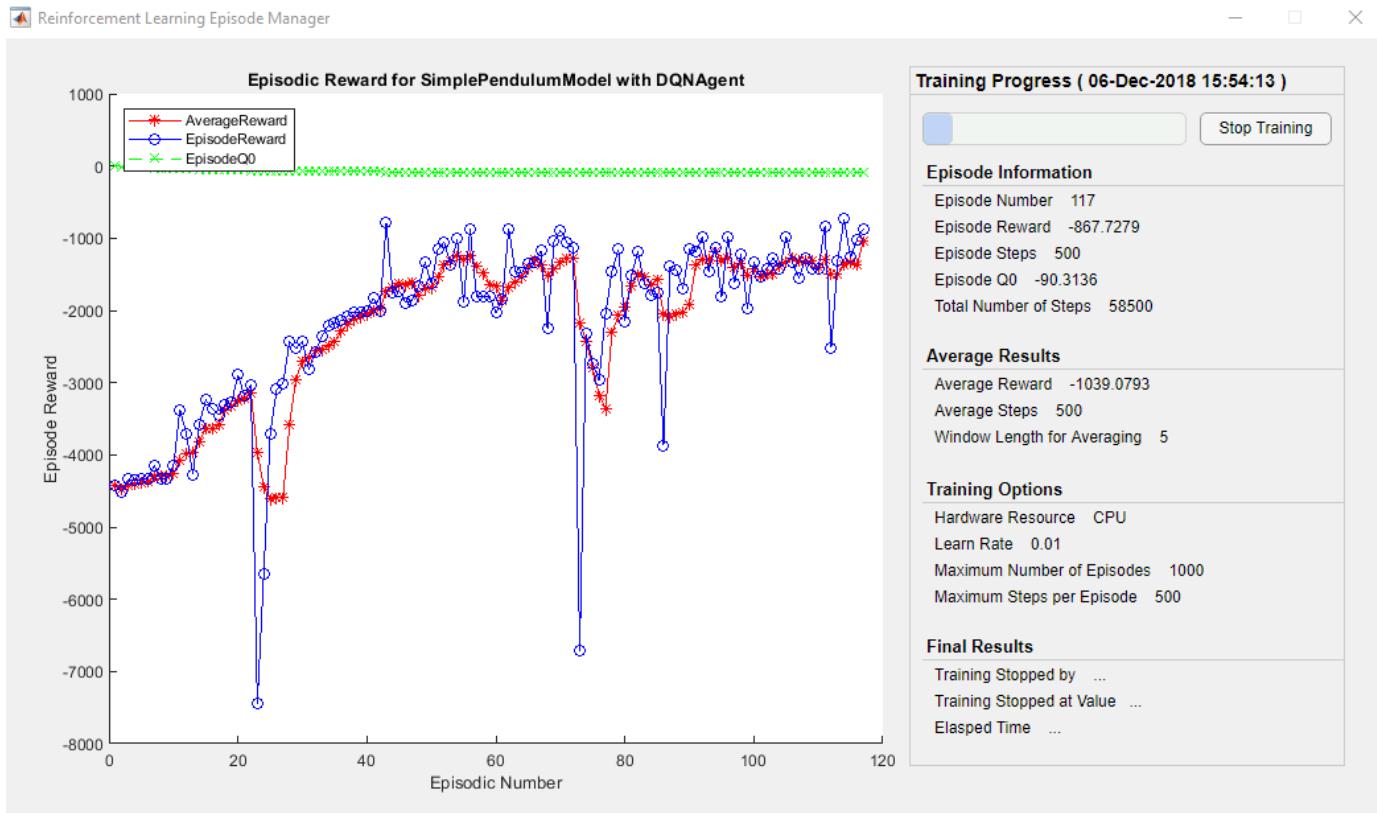
For more information, see `rlTrainingOptions`.

```
trainingOptions = rlTrainingOptions(...  
    'MaxEpisodes',1000,...  
    'MaxStepsPerEpisode',500,...  
    'ScoreAveragingWindowLength',5,...  
    'Verbose',false,...  
    'Plots','training-progress',...  
    'StopTrainingCriteria','AverageReward',...  
    'StopTrainingValue',-1100,...  
    'SaveAgentCriteria','EpisodeReward',...  
    'SaveAgentValue',-1100);
```

Train the agent using the `train` function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;  
  
if doTraining
```

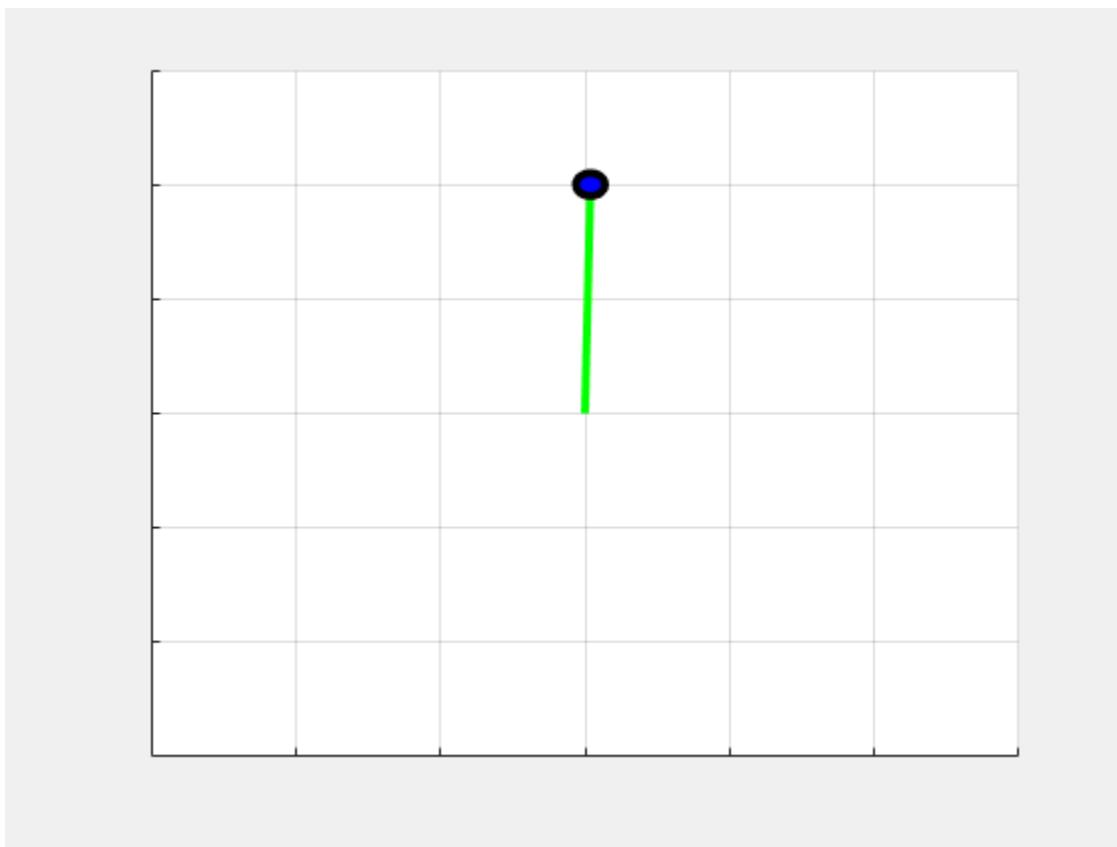
```
% Train the agent.
trainingStats = train(agent,env,trainingOptions);
else
    % Load pretrained agent for the example.
    load('SimulinkPendulumDQN.mat','agent');
end
```



Simulate DQN Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



See Also

[rlDQNAgent](#)

More About

- “Deep Q-Network Agents” on page 4-8
- “Create Simulink Environments for Reinforcement Learning” on page 2-5

Train DDPG Agent to Swing Up and Balance Pendulum

This example shows how to train a deep deterministic policy gradient (DDPG) agent to swing up and balance a pendulum modeled in Simulink®.

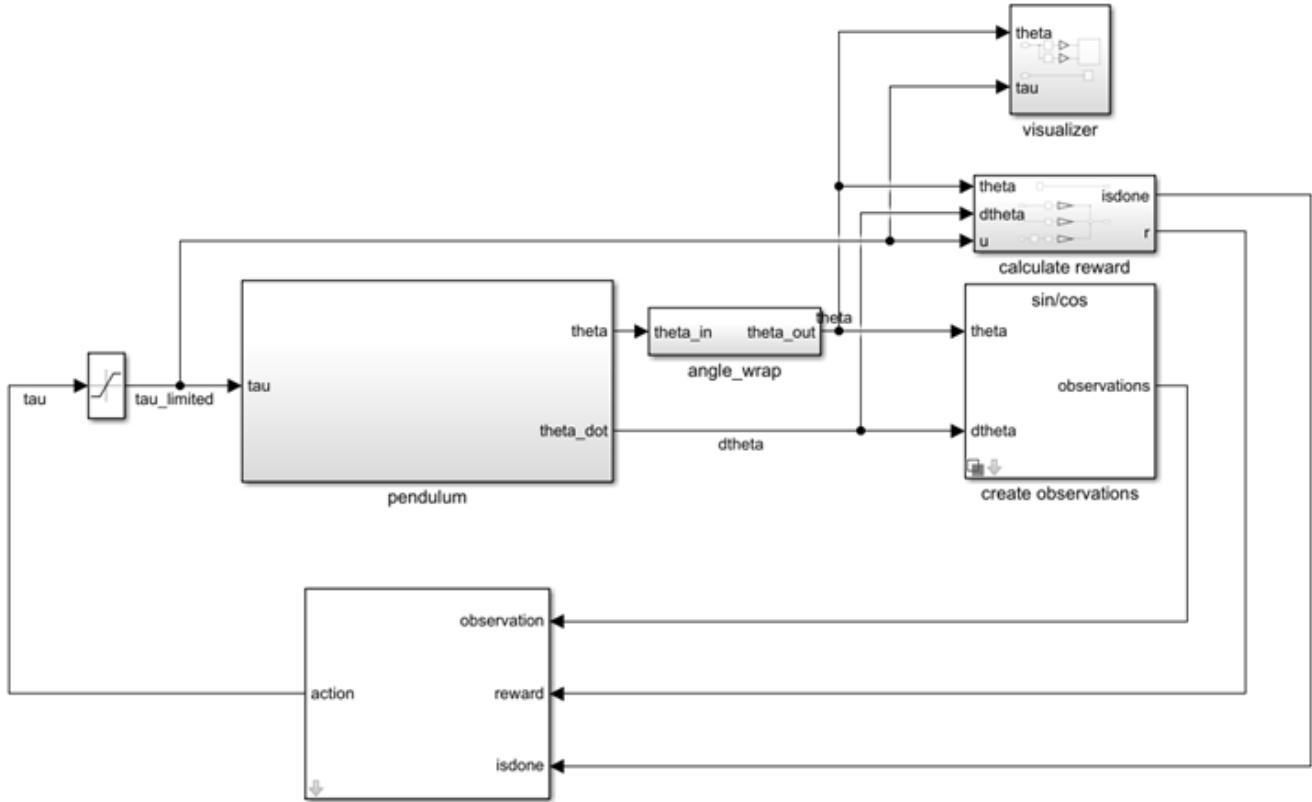
For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14. For an example that trains a DDPG agent in MATLAB®, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Pendulum Swing-up Model

The reinforcement learning environment for this example is a simple frictionless pendulum that is initially hanging in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

Open the model.

```
mdl = 'rlSimplePendulumModel';
open_system(mdl)
```



For this model:

- The upward balanced pendulum position is $\theta = 0$ radians, and the downward hanging position is $\theta = \pi$ radians.
- The torque action signal from the agent to the environment is from -2 to 2 Nm.

- The observations from the environment are the sine of the pendulum angle, the cosine of the pendulum angle, and the pendulum angle derivative.
- The reward r_t , provided at every time step, is:

$$r_t = -(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001u_{t-1}^2)$$

where:

- θ_t is the angle of displacement from the upright position
- $\dot{\theta}_t$ is the derivative of the displacement angle
- u_{t-1} is the control effort from the previous time step.

For more information on this model, see “Load Predefined Simulink Environments” on page 2-22.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv('SimplePendulumModel-Continuous')

env =
  SimulinkEnvWithAgent with properties:

    Model: "rlSimplePendulumModel"
    AgentBlock: "rlSimplePendulumModel/RL Agent"
    ResetFcn: []
    UseFastRestart: 'on'
```

The interface has a continuous action space where the agent can apply possible torque values from -2 to 2 Nm to the pendulum.

Set the observations of the environment to be the sine of the pendulum angle, the cosine of the pendulum angle, and the pendulum angle derivative.

```
numObs = 3;
set_param('rlSimplePendulumModel/create_observations','ThetaObservationHandling','sincos');
```

To define the initial condition of the pendulum as hanging downward, specify an environment reset function using an anonymous function handle. This reset function sets the model workspace variable `theta0` to `pi`.

```
env.ResetFcn = @(in)setVariable(in,'theta0',pi,'Workspace',mdl);
```

Specify the simulation time `Tf` and the agent sample time `Ts` in seconds

```
Ts = 0.05;
Tf = 20;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

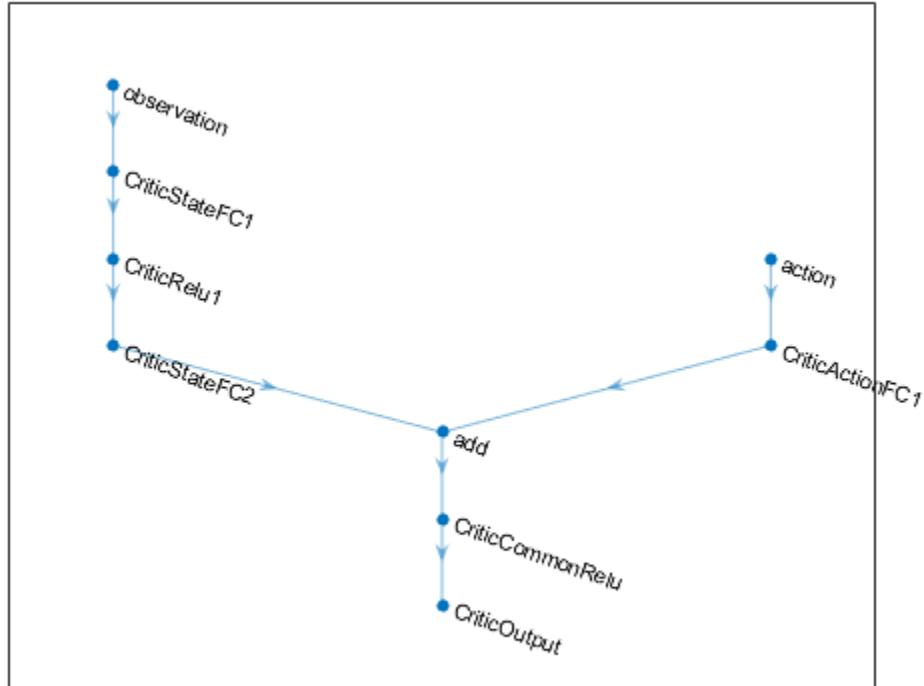
```
statePath = [
    imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(400, 'Name', 'CriticStateFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(300, 'Name', 'CriticStateFC2')];
actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(300, 'Name', 'CriticActionFC1', 'BiasLearnRateFactor', 0)];
commonPath = [
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'CriticCommonRelu')
    fullyConnectedLayer(1, 'Name', 'CriticOutput')];

criticNetwork = layerGraph();
criticNetwork = addLayers(criticNetwork, statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = addLayers(criticNetwork, commonPath);

criticNetwork = connectLayers(criticNetwork, 'CriticStateFC2', 'add/in1');
criticNetwork = connectLayers(criticNetwork, 'CriticActionFC1', 'add/in2');
```

View the critic network configuration.

```
figure
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
obsInfo = getObservationInfo(env);
actInfo = getActionInfo(env);
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,'Observation',{'observation'},'Action','action');
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input, the observation, and one output, the action.

Construct the actor similarly to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [
    imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(400, 'Name', 'ActorFC1')
    reluLayer('Name', 'ActorRelu1')
    fullyConnectedLayer(300, 'Name', 'ActorFC2')
    reluLayer('Name', 'ActorRelu2')
    fullyConnectedLayer(1, 'Name', 'ActorFC3')
    tanhLayer('Name', 'ActorTanh')
```

```

scalingLayer('Name','ActorScaling','Scale',max(actInfo.UpperLimit))];

actorOpts = rlRepresentationOptions('LearnRate',1e-04,'GradientThreshold',1);

actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,'Observation',{ 'observat

```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```

agentOpts = rlDDPGAgentOptions(...  

    'SampleTime',Ts,...  

    'TargetSmoothFactor',1e-3,...  

    'ExperienceBufferLength',1e6,...  

    'DiscountFactor',0.99,...  

    'MiniBatchSize',128);  

agentOpts.NoiseOptions.Variance = 0.6;  

agentOpts.NoiseOptions.VarianceDecayRate = 1e-5;

```

Then, create the DDPG agent using the specified actor representation, critic representation and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run training for at most **50000** episodes, with each episode lasting at most `ceil(Tf/Ts)` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than **-740** over five consecutive episodes. At this point, the agent can quickly balance the pendulum in the upright position using minimal control effort.
- Save a copy of the agent for each episode where the cumulative reward is greater than **-740**.

For more information, see `rlTrainingOptions`.

```

maxepisodes = 5000;  

maxsteps = ceil(Tf/Ts);  

trainOpts = rlTrainingOptions(...  

    'MaxEpisodes',maxepisodes,...  

    'MaxStepsPerEpisode',maxsteps,...  

    'ScoreAveragingWindowLength',5,...  

    'Verbose',false,...  

    'Plots','training-progress',...  

    'StopTrainingCriteria','AverageReward',...  

    'StopTrainingValue',-740,...  

    'SaveAgentCriteria','EpisodeReward',...  

    'SaveAgentValue',-740);

```

Train the agent using the `train` function. Training this agent is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

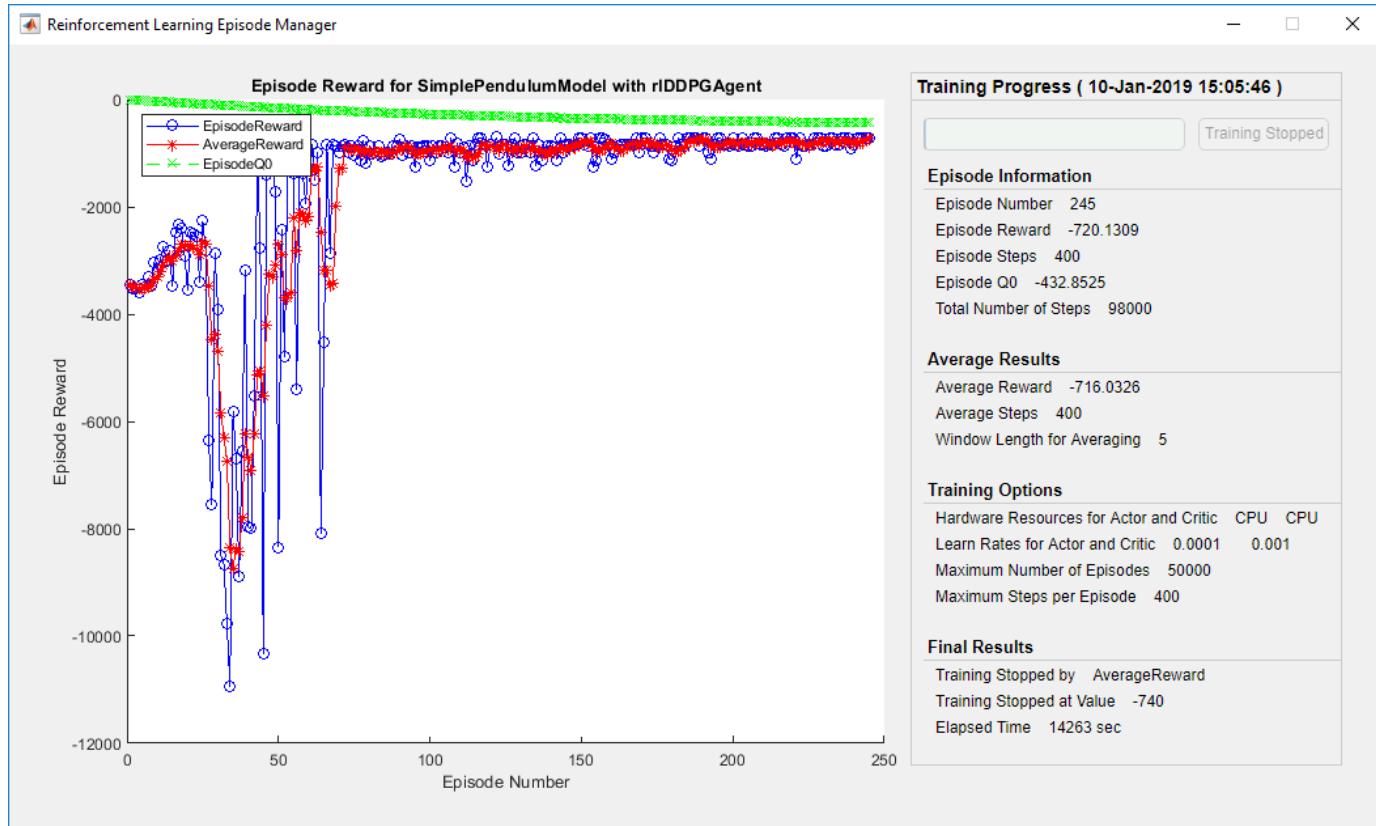
```

doTraining = false;  

if doTraining

```

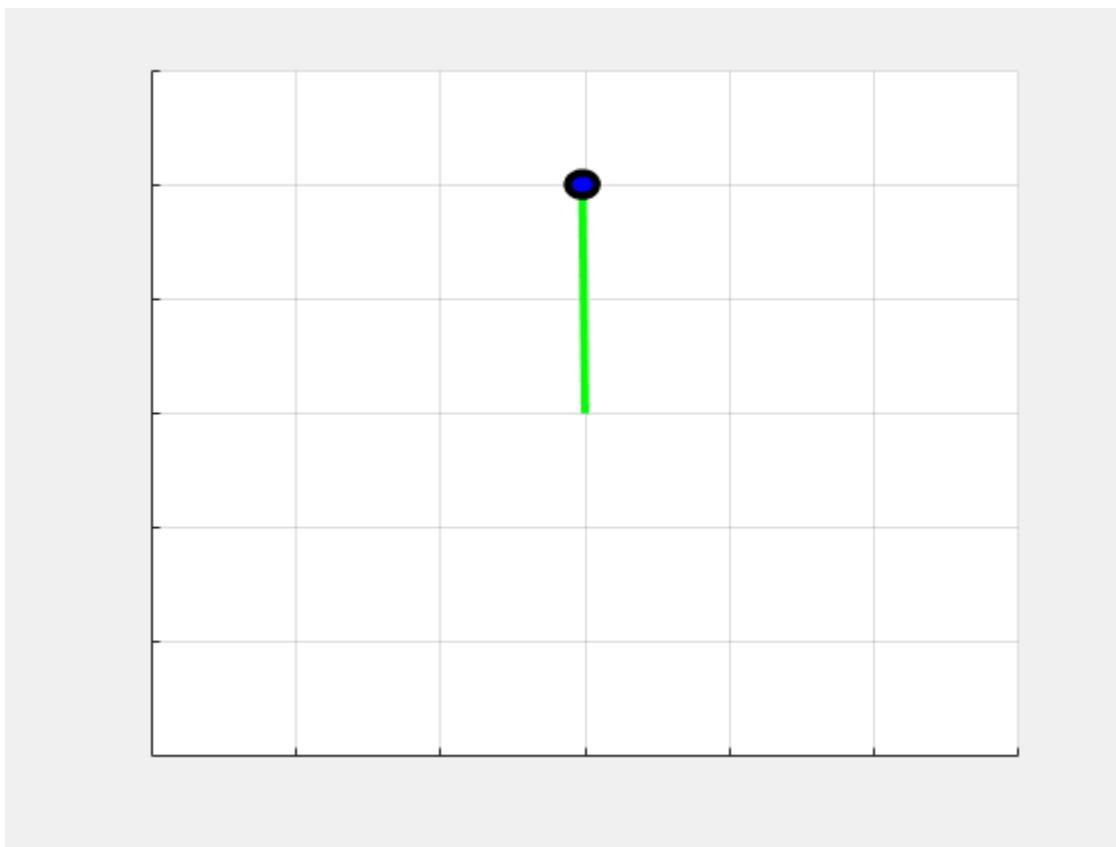
```
% Train the agent.
trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('SimulinkPendulumDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



See Also

[rlDDPGAgent](#) | [rlSimulinkEnv](#) | [train](#)

More About

- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Deep Deterministic Policy Gradient Agents” on page 4-14

Train DDPG Agent to Swing Up and Balance Cart-Pole System

This example shows how to train a deep deterministic policy gradient (DDPG) agent to swing up and balance a cart-pole system modeled in Simscape™ Multibody™.

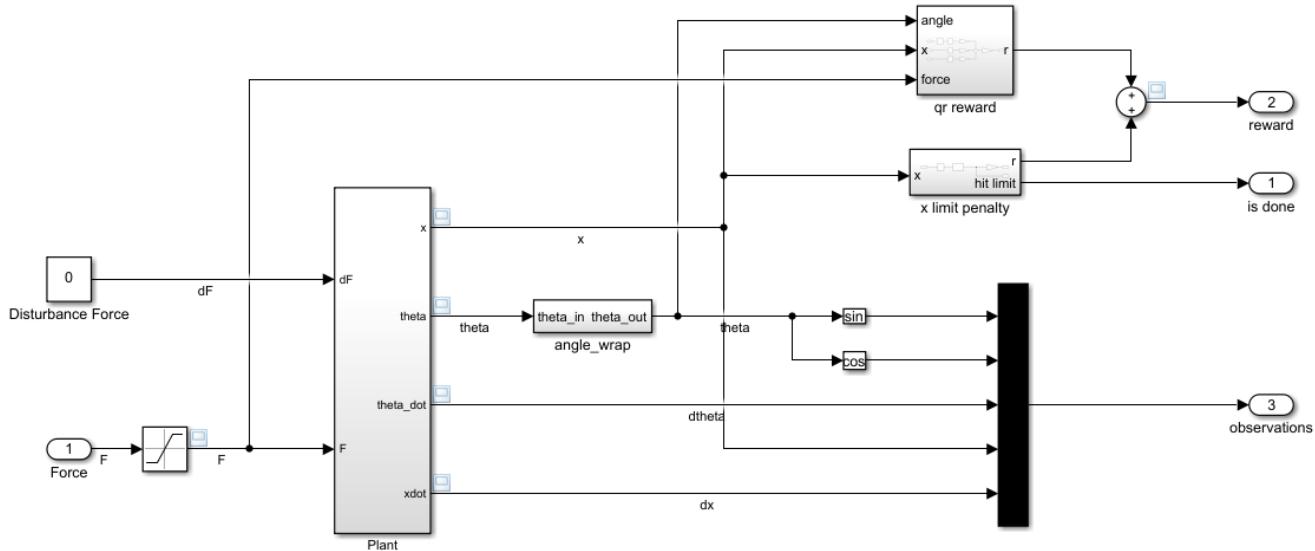
For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14. For an example that trains a DDPG agent in MATLAB®, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Cart-Pole Simscape Model

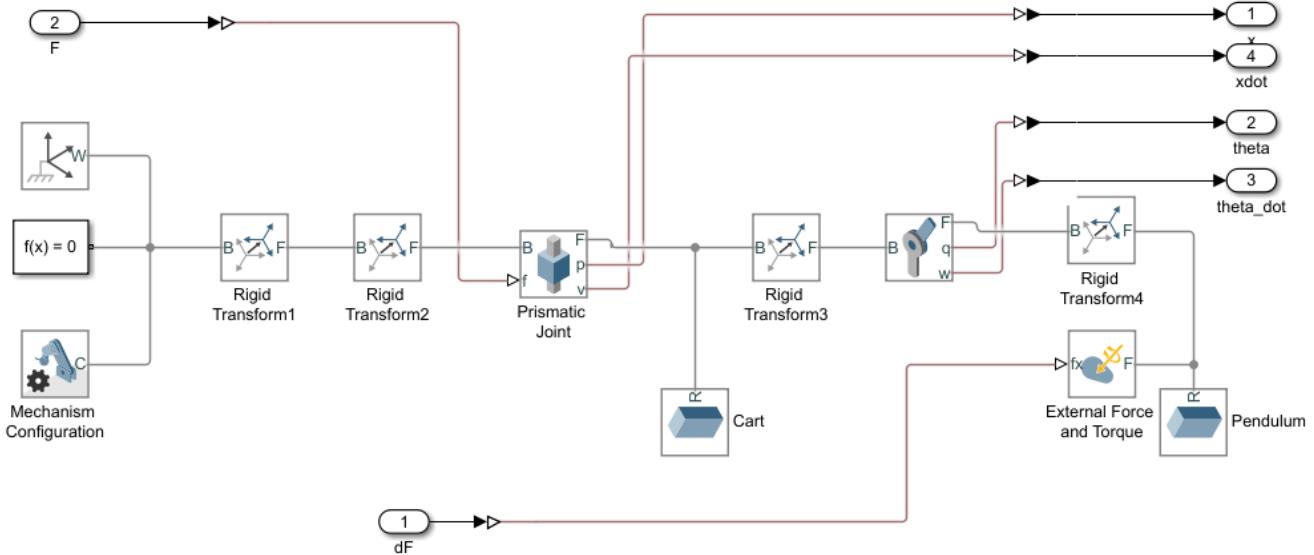
The reinforcement learning environment for this example is a pole attached to an unactuated joint on a cart, which moves along a frictionless track. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

Open the model.

```
mdl = 'rlCartPoleSimscapeModel';
open_system(mdl)
```



The cart-pole system is modeled using Simscape Multibody.



For this model:

- The upward balanced pendulum position is 0 radians, and the downward hanging position is pi radians
- The force action signal from the agent to the environment is from - 15 to 15 N
- The observations from the environment are the position and velocity of the cart, and the sine, cosine, and derivative of the pendulum angle
- The episode terminates if the cart moves more than 3.5 m from the original position
- The reward r_t , provided at every timestep, is:

$$r_t = -0.1(5\theta_t^2 + x_t^2 + 0.05u_{t-1}^2) - 100B$$

where:

- θ_t is the angle of displacement from the upright position of the pendulum
- x_t is the position displacement from the center position of the cart
- u_{t-1} is the control effort from the previous time step
- B is a flag (1 or 0) that indicates whether the cart is out of bounds

For more information on this model, see “Load Predefined Simulink Environments” on page 2-22.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv('CartPoleSimscapeModel-Continuous')
env =
SimulinkEnvWithAgent with properties:
```

```
Model: "rlCartPoleSimscapeModel"
AgentBlock: "rlCartPoleSimscapeModel/RL Agent"
ResetFcn: []
UseFastRestart: 'on'
```

The interface has a continuous action space where the agent can apply possible torque values from -15 to 15 N to the pendulum.

Obtain the observation and action information from the environment interface.

```
obsInfo = getObservationInfo(env);
numObservations = obsInfo.Dimension(1);
actInfo = getActionInfo(env);
```

Specify the simulation time T_f and the agent sample time T_s in seconds

```
Ts = 0.02;
Tf = 25;
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs (the state and action) and one output. The input size of action path is [1 1 1] since the agent can apply action as one force value to the environment. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
statePath = [
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(128, 'Name', 'CriticStateFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(200, 'Name', 'CriticStateFC2')];

actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(200, 'Name', 'CriticActionFC1', 'BiasLearnRateFactor', 0)]];

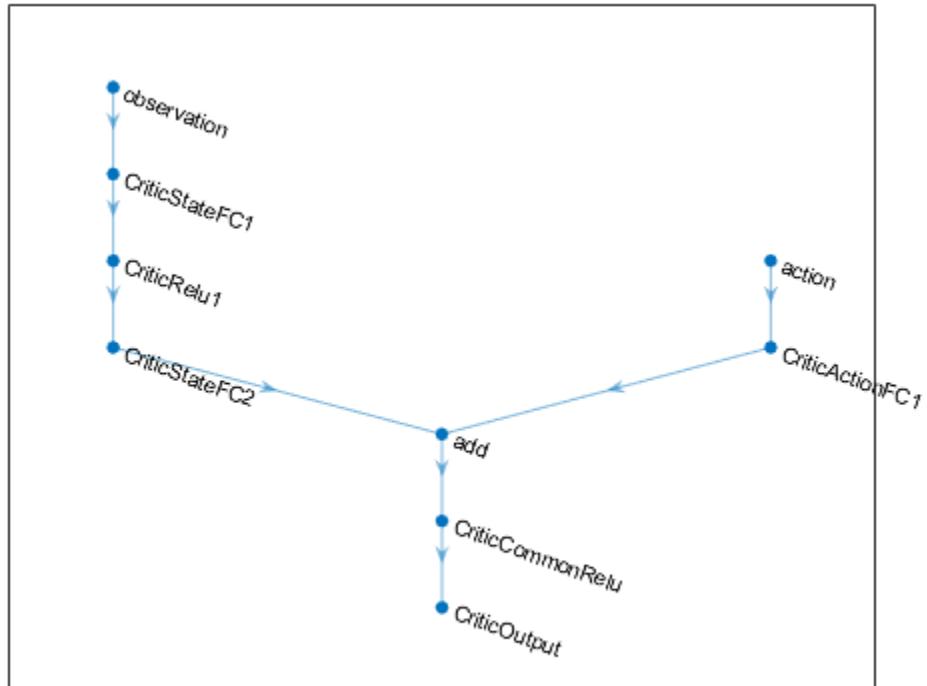
commonPath = [
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'CriticCommonRelu')
    fullyConnectedLayer(1, 'Name', 'CriticOutput')];

criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = addLayers(criticNetwork, commonPath);

criticNetwork = connectLayers(criticNetwork, 'CriticStateFC2', 'add/in1');
criticNetwork = connectLayers(criticNetwork, 'CriticActionFC1', 'add/in2');
```

View the critic network configuration.

```
figure
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation information for the critic, which you already obtained from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,...  
    'Observation',{'observation'},'Action',{'action'},criticOptions);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input (the observation) and one output (the action).

Construct the actor in a similar manner to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [  
    imageInputLayer([numObservations 1 1],'Normalization','none','Name','observation')  
    fullyConnectedLayer(128,'Name','ActorFC1')  
    reluLayer('Name','ActorRelu1')  
    fullyConnectedLayer(200,'Name','ActorFC2')  
    reluLayer('Name','ActorRelu2')  
    fullyConnectedLayer(1,'Name','ActorFC3')  
    tanhLayer('Name','ActorTanh1')  
    scalingLayer('Name','ActorScaling','Scale',max(actInfo.UpperLimit))];
```

```
actorOptions = rlRepresentationOptions('LearnRate',5e-04,'GradientThreshold',1);  
  
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,...  
    'Observation',{'observation'},'Action',{'ActorScaling'},actorOptions);
```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```
agentOptions = rlDDPGAgentOptions(...  
    'SampleTime',Ts,...  
    'TargetSmoothFactor',1e-3,...  
    'ExperienceBufferLength',1e6,...  
    'MiniBatchSize',128);  
agentOptions.NoiseOptions.Variance = 0.4;  
agentOptions.NoiseOptions.VarianceDecayRate = 1e-5;
```

Then, create the agent using the specified actor representation, critic representation and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training episode for at most 2000 episodes, with each episode lasting at most `ceil(Tf/Ts)` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than -400 over five consecutive episodes. At this point, the agent can quickly balance the pendulum in the upright position using minimal control effort.
- Save a copy of the agent for each episode where the cumulative reward is greater than -400.

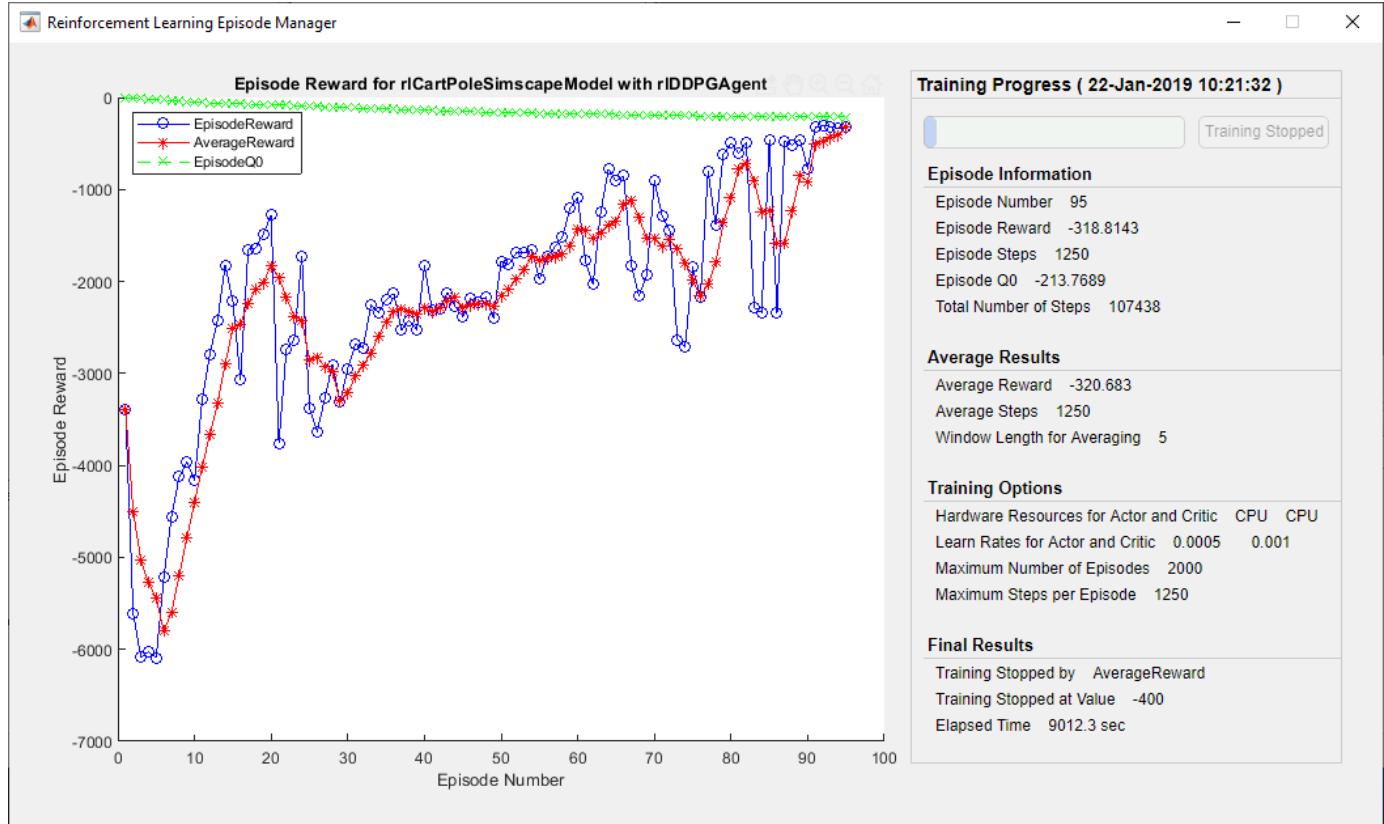
For more information, see `rlTrainingOptions`.

```
maxepisodes = 2000;  
maxsteps = ceil(Tf/Ts);  
trainingOptions = rlTrainingOptions(...  
    'MaxEpisodes',maxepisodes,...  
    'MaxStepsPerEpisode',maxsteps,...  
    'ScoreAveragingWindowLength',5,...  
    'Verbose',false,...  
    'Plots','training-progress',...  
    'StopTrainingCriteria','AverageReward',...  
    'StopTrainingValue',-400,...  
    'SaveAgentCriteria','EpisodeReward',...  
    'SaveAgentValue',-400);
```

Train the agent using the `train` function. This process is computationally intensive and takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;  
  
if doTraining
```

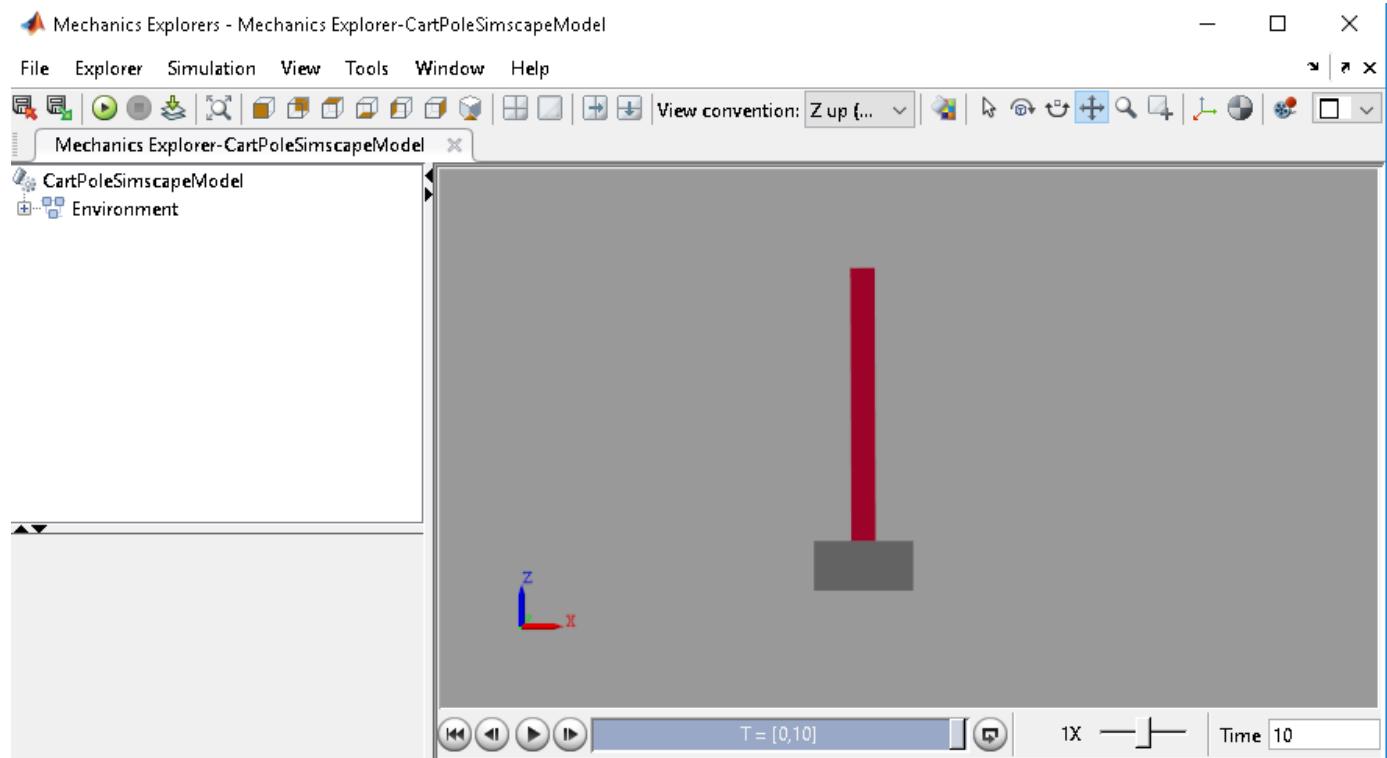
```
% Train the agent.
trainingStats = train(agent,env,trainingOptions);
else
    % Load pretrained agent for the example.
    load('SimscapeCartPoleDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the cart-pole environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
bdclose(mdl)
```

See Also

[rlDDPGAgent](#) | [rlSimulinkEnv](#) | [train](#)

More About

- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Deep Deterministic Policy Gradient Agents” on page 4-14

Train DDPG Agent to Swing Up and Balance Pendulum with Bus Signal

Starting with a simple frictionless pendulum Simulink® model, this example shows how to convert it to a reinforcement learning environment interface and trains a deep deterministic policy gradient (DDPG) agent in this environment.

For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14. For an example that trains a DDPG agent in MATLAB®, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Pendulum Swing-Up Model with Bus

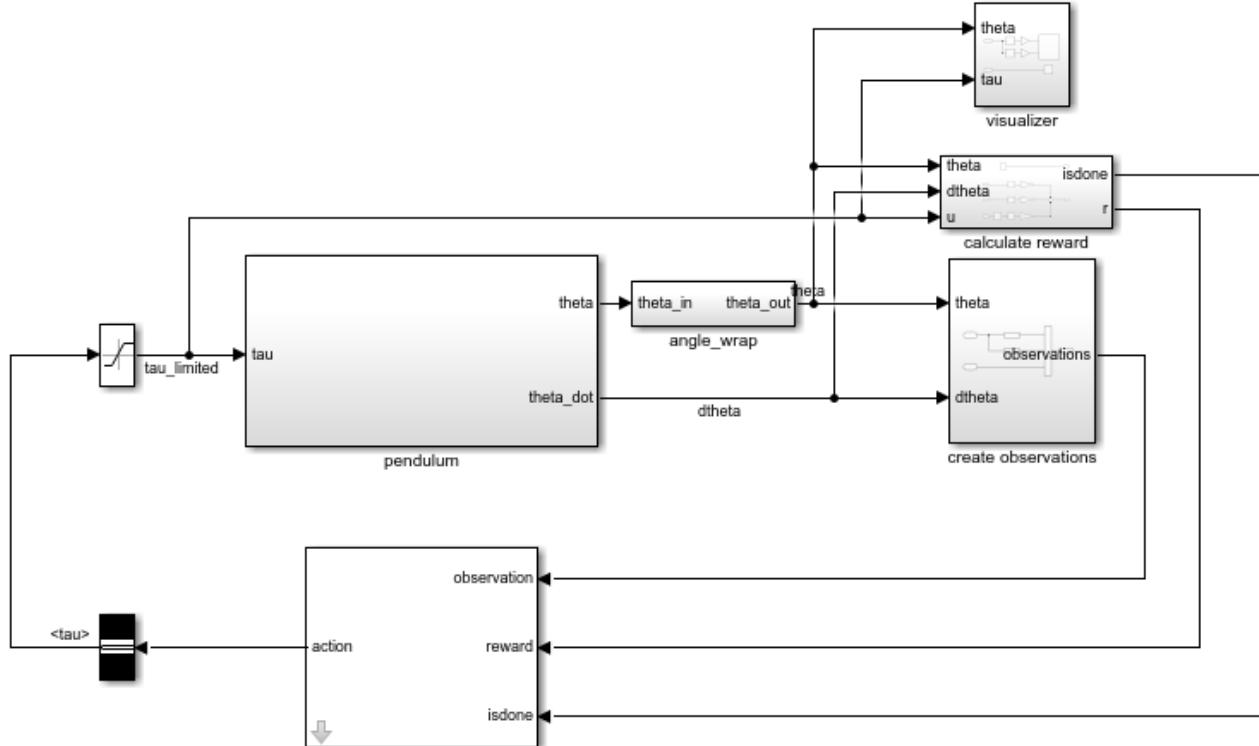
The starting model for this example is a simple frictionless pendulum. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

Fix the random generator seed for reproducibility.

```
rng(0)
```

Open the model.

```
mdl = 'rlSimplePendulumModelBus';
open_system(mdl)
```



For this model:

- The upward balanced pendulum position is θ radians, and the downward hanging position is π radians.
- The torque action signal from the agent to the environment is from -2 to 2 Nm.
- The observations from the environment are the sine of the pendulum angle, the cosine of the pendulum angle, and the pendulum angle derivative.
- Both the observation and action signals are Simulink buses.
- The reward r_t , provided at every time step, is:

$$r_t = -(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001u_{t-1}^2)$$

where:

- θ_t is the angle of displacement from the upright position
- $\dot{\theta}_t$ is the derivative of the displacement angle
- u_{t-1} is the control effort from the previous time step.

The model used in this example is similar to the simple pendulum model described in “Load Predefined Simulink Environments” on page 2-22. The difference is that the model in this example uses Simulink buses for the action and observation signals.

Create Environment Interface with Bus

The environment interface from a Simulink model is created using `rlSimulinkEnv`, which requires the name of the Simulink model, the path the agent block, and observation and action reinforcement learning data specifications. For models that use bus signals for actions or observations, you can create the corresponding specifications using the `bus2RLSpec` function.

Specify the path to the agent block.

```
agentBlk = 'rlSimplePendulumModelBus/RL_Agent';
```

Create the observation Bus object.

```
obsBus = Simulink.Bus();
obs(1) = Simulink.BusElement;
obs(1).Name = 'sin_theta';
obs(2) = Simulink.BusElement;
obs(2).Name = 'cos_theta';
obs(3) = Simulink.BusElement;
obs(3).Name = 'dtheta';
obsBus.Elements = obs;
```

Create the action Bus object.

```
actBus = Simulink.Bus();
act(1) = Simulink.BusElement;
act(1).Name = 'tau';
act(1).Min = -2;
act(1).Max = 2;
actBus.Elements = act;
```

Create the action and observation specification objects using the Simulink buses.

```
obsInfo = bus2RLSpec('obsBus','Model',mdl);
actInfo = bus2RLSpec('actBus','Model',mdl);
```

Create the reinforcement learning environment for the pendulum model.

```
env = rlSimulinkEnv(mdl,agentBlk,obsInfo,actInfo);
```

To define the initial condition of the pendulum as hanging downward, specify an environment reset function using an anonymous function handle. This reset function sets the model workspace variable `theta0` to `pi`.

```
env.ResetFcn = @(in)setVariable(in,'theta0',pi,'Workspace',mdl);
```

Specify the simulation time `Tf` and the agent sample time `Ts` in seconds

```
Ts = 0.05;
Tf = 20;
```

Create DDPG agent

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with three inputs, the three observations, and one output, the action. The three observations can be combined using a `concatenationLayer`.

For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

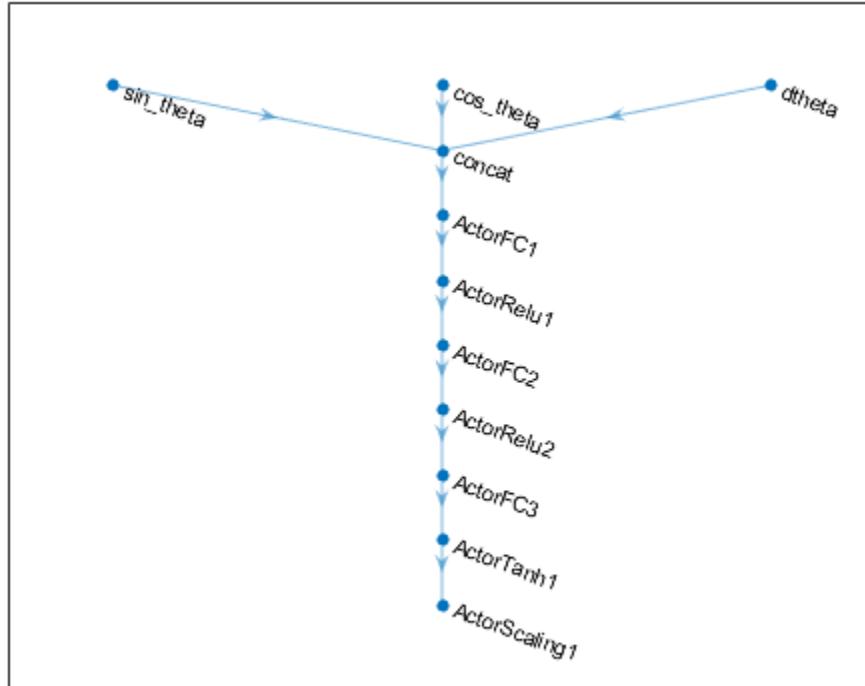
```
sinThetaInput = imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'sin_theta');
cosThetaInput = imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'cos_theta');
dThetaInput = imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'dtheta');
commonPath = [
    concatenationLayer(1,3, 'Name', 'concat')
    fullyConnectedLayer(400, 'Name', 'ActorFC1')
    reluLayer('Name', 'ActorRelu1')
    fullyConnectedLayer(300, 'Name', 'ActorFC2')
    reluLayer('Name', 'ActorRelu2')
    fullyConnectedLayer(1, 'Name', 'ActorFC3')
    tanhLayer('Name', 'ActorTanh1')
    scalingLayer('Name', 'ActorScaling1', 'Scale', max(actInfo.UpperLimit))];

actorNetwork = layerGraph(sinThetaInput);
actorNetwork = addLayers(actorNetwork,cosThetaInput);
actorNetwork = addLayers(actorNetwork,dThetaInput);
actorNetwork = addLayers(actorNetwork,commonPath);

actorNetwork = connectLayers(actorNetwork, 'sin_theta', 'concat/in1');
actorNetwork = connectLayers(actorNetwork, 'cos_theta', 'concat/in2');
actorNetwork = connectLayers(actorNetwork, 'dtheta', 'concat/in3');
```

View the actor network configuration.

```
figure
plot(actorNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
actorOptions = rlRepresentationOptions('LearnRate',1e-4,'GradientThreshold',1);
```

Create the actor representation using the specified deep neural network and options. You must also specify the action and observation info for the actor, which you obtained from the environment interface. For more information, see `rlDeterministicActorRepresentation`.

```
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,...  
'Observation',{'sin_theta','cos_theta','dtheta'},'Action',{'ActorScaling1'},actorOptions);
```

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the observation and action, and one output, the state action value.

Construct the critic similarly to the actor. For more information, see `rlQValueRepresentation`.

```
statePath = [  
    concatenationLayer(1,3,'Name','concat')  
    fullyConnectedLayer(400,'Name','CriticStateFC1')  
    reluLayer('Name','CriticRelu1')  
    fullyConnectedLayer(300,'Name','CriticStateFC2')];  
  
actionPath = [  
    imageInputLayer([1 1 1],'Normalization','none','Name','action')  
    fullyConnectedLayer(300,'Name','CriticActionFC1','BiasLearnRateFactor', 0)];
```

```

commonPath = [
    additionLayer(2,'Name','add')
    reluLayer('Name','CriticCommonRelu')
    fullyConnectedLayer(1,'Name','CriticOutput')];

criticNetwork = layerGraph(sinThetaInput);
criticNetwork = addLayers(criticNetwork,cosThetaInput);
criticNetwork = addLayers(criticNetwork,dThetaInput);
criticNetwork = addLayers(criticNetwork,actionPath);
criticNetwork = addLayers(criticNetwork,statePath);
criticNetwork = addLayers(criticNetwork,commonPath);

criticNetwork = connectLayers(criticNetwork,'sin_theta','concat/in1');
criticNetwork = connectLayers(criticNetwork,'cos_theta','concat/in2');
criticNetwork = connectLayers(criticNetwork,'dtheta','concat/in3');
criticNetwork = connectLayers(criticNetwork,'CriticStateFC2','add/in1');
criticNetwork = connectLayers(criticNetwork,'CriticActionFC1','add/in2');

criticOpts = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);

critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,...  

    'Observation',{ 'sin_theta','cos_theta','dtheta'},'Action',{ 'action'},cr...

```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```

agentOpts = rlDDPGAgentOptions(...  

    'SampleTime',Ts,...  

    'TargetSmoothFactor',1e-3,...  

    'ExperienceBufferLength',1e6,...  

    'DiscountFactor',0.99,...  

    'MiniBatchSize',128);  

agentOpts.NoiseOptions.Variance = 0.6;  

agentOpts.NoiseOptions.VarianceDecayRate = 1e-5;

```

Then, create the DDPG agent using the specified actor representation, critic representation and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training for at most 50000 episodes, with each episode lasting at most `ceil(Tf/Ts)` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than -740 over five consecutive episodes. At this point, the agent can quickly balance the pendulum in the upright position using minimal control effort.
- Save a copy of the agent for each episode where the cumulative reward is greater than -740.

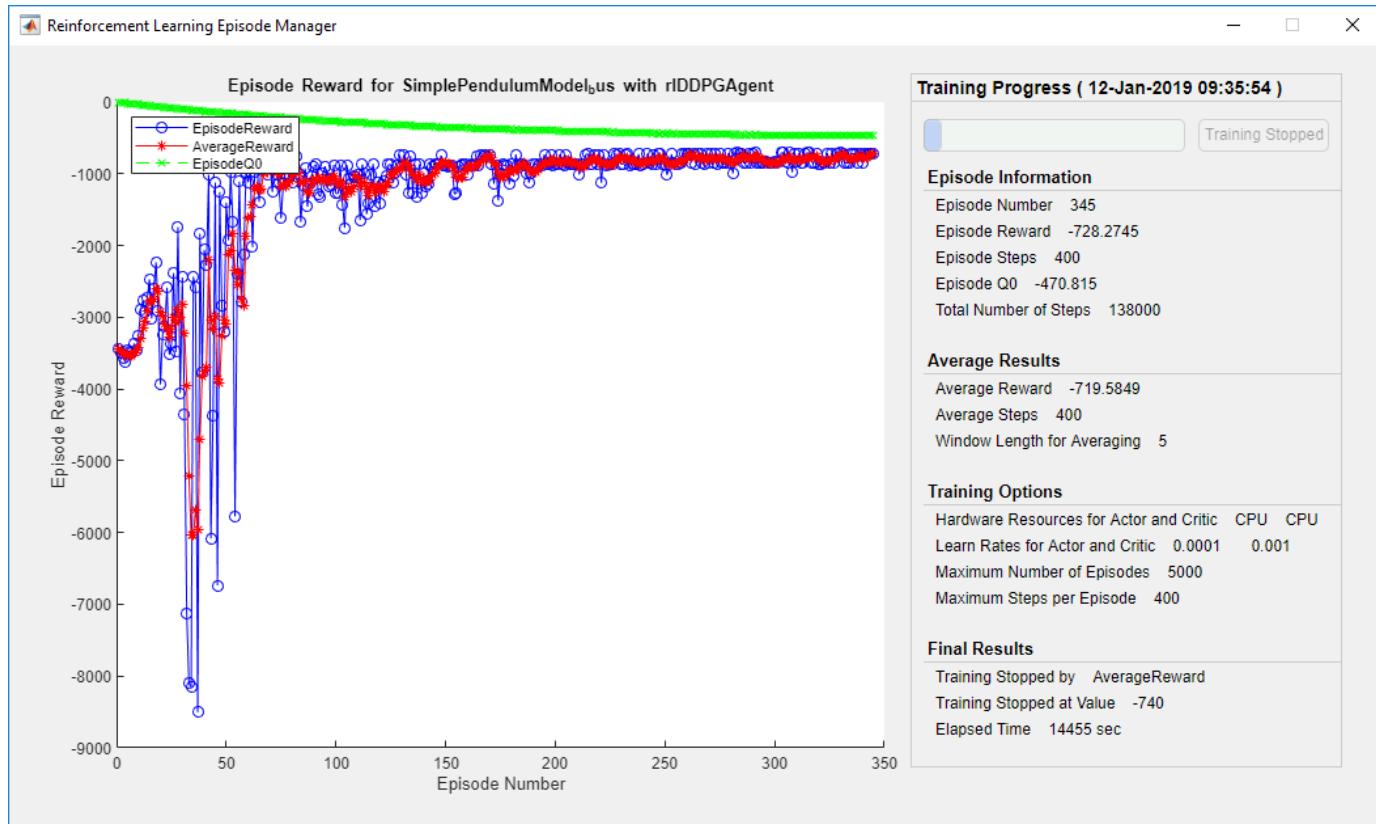
For more information, see `rlTrainingOptions`.

```
maxepisodes = 5000;  
maxsteps = ceil(Tf/Ts);
```

```
trainOpts = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes,...
    'MaxStepsPerEpisode',maxsteps,...
    'ScoreAveragingWindowLength',5,...
    'Verbose',false,...
    'Plots','training-progress',...
    'StopTrainingCriteria','AverageReward',...
    'StopTrainingValue',-740);
```

Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

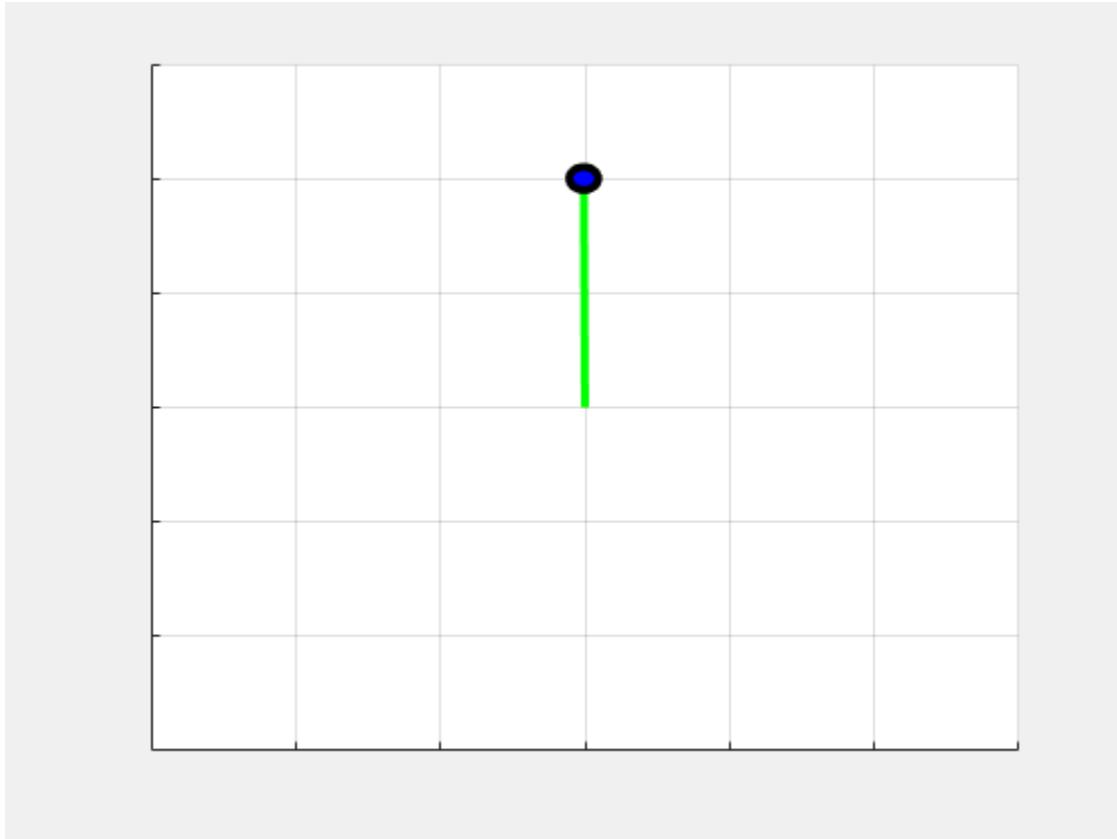
```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('SimulinkPendBusDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);  
experience = sim(env,agent,simOptions);
```



See Also

[bus2RLSpec](#) | [rlDDPGAgent](#) | [rlSimulinkEnv](#) | [train](#)

More About

- “Create Simulink Environments for Reinforcement Learning” on page 2-5
- “Deep Deterministic Policy Gradient Agents” on page 4-14

Train DDPG Agent to Swing Up and Balance Pendulum with Image Observation

This example shows how to train a deep deterministic policy gradient (DDPG) agent to swing up and balance a pendulum with an image observation modeled in MATLAB®.

For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14.

Simple Pendulum with Image MATLAB Environment

The reinforcement learning environment for this example is a simple frictionless pendulum that is initially hanging in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort.

For this environment:

- The upward balanced pendulum position is θ radians, and the downward hanging position is π radians
- The torque action signal from the agent to the environment is from -2 to 2 Nm
- The observations from the environment are an image indicating the location of the pendulum's mass and the pendulum angular velocity.
- The reward r_t , provided at every time step, is:

$$r_t = -\left(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001u_{t-1}^2\right)$$

where:

- θ_t is the angle of displacement from the upright position
- $\dot{\theta}_t$ is the derivative of the displacement angle
- u_{t-1} is the control effort from the previous time step

For more information on this model, see “Load Predefined Control System Environments” on page 2-15.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv('SimplePendulumWithImage-Continuous')
env =
  SimplePendulumWithImageContinuousAction with properties:
    Mass: 1
    RodLength: 1
    RodInertia: 0
    Gravity: 9.8100
    DampingRatio: 0
    MaximumTorque: 2
    Ts: 0.0500
    State: [2x1 double]
```

```

Q: [2x2 double]
R: 1.0000e-03

```

The interface has a continuous action space where the agent can apply a torque between -2 to 2 Nm.

Obtain the observation and action specification from the environment interface.

```

obsInfo = getObservationInfo(env);
actInfo = getActionInfo(env);

```

Fix the random generator seed for reproducibility.

```

rng(0)

```

Create DDPG Agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep convolutional neural network (CNN) with three inputs (the image, angular velocity, and action) and one output. For more information on creating representations, see “Create Policy and Value Function Representations” on page 3-2.

```

hiddenLayerSize1 = 400;
hiddenLayerSize2 = 300;

imgPath = [
    imageInputLayer(obsInfo(1).Dimension, 'Normalization', 'none', 'Name', obsInfo(1).Name)
    convolution2dLayer(10, 2, 'Name', 'conv1', 'Stride', 5, 'Padding', 0)
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(2, 'Name', 'fc1')
    concatenationLayer(3, 2, 'Name', 'cat1')
    fullyConnectedLayer(hiddenLayerSize1, 'Name', 'fc2')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(hiddenLayerSize2, 'Name', 'fc3')
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(1, 'Name', 'fc4')
];
dthetaPath = [
    imageInputLayer(obsInfo(2).Dimension, 'Normalization', 'none', 'Name', obsInfo(2).Name)
    fullyConnectedLayer(1, 'Name', 'fc5', 'BiasLearnRateFactor', 0, 'Bias', 0)
];
actPath =[ 
    imageInputLayer(actInfo(1).Dimension, 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(hiddenLayerSize2, 'Name', 'fc6', 'BiasLearnRateFactor', 0, 'Bias', zeros(hiddenLayerSize2))
];

criticNetwork = layerGraph(imgPath);
criticNetwork = addLayers(criticNetwork, dthetaPath);
criticNetwork = addLayers(criticNetwork, actPath);
criticNetwork = connectLayers(criticNetwork, 'fc5', 'cat1/in2');
criticNetwork = connectLayers(criticNetwork, 'fc6', 'add/in2');

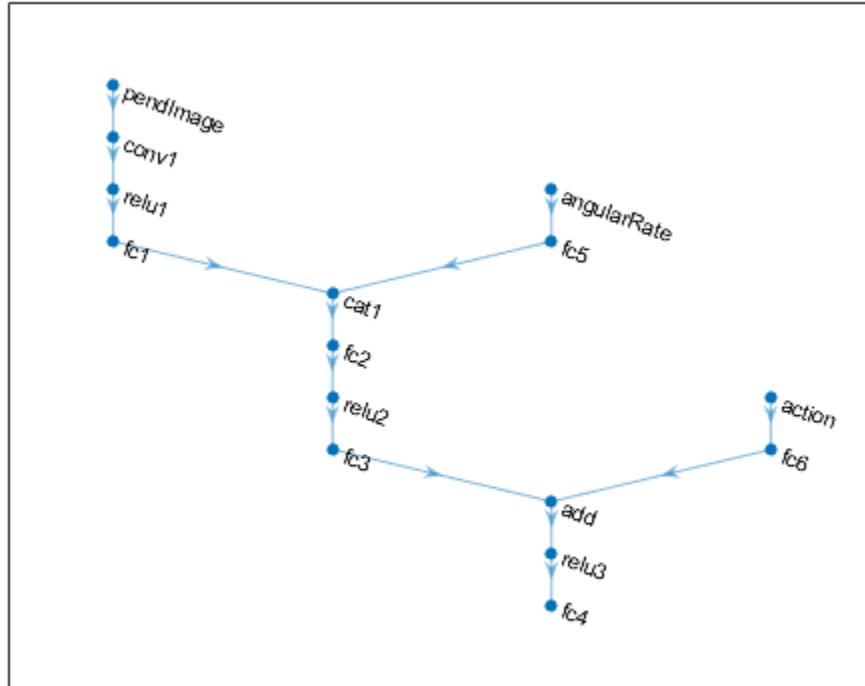
```

View the critic network configuration.

```

figure
plot(criticNetwork)

```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);
```

Uncomment the following line to use the GPU to accelerate training of the critic CNN.

```
% criticOptions.UseDevice = 'gpu';
```

Create the critic representation using the specified neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,obsInfo,actInfo,...  
    'Observation',{'pendImage','angularRate'},'Action',{'action'},criticOptions);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep convolutional neural network (CNN) with two inputs (the image and angular velocity) and one output (the action).

Construct the actor in a similar manner to the critic.

```
imgPath = [  
    imageInputLayer(obsInfo(1).Dimension,'Normalization','none','Name',obsInfo(1).Name)  
    convolution2dLayer(10,2,'Name','conv1','Stride',5,'Padding',0)  
    reluLayer('Name','relu1')  
    fullyConnectedLayer(2,'Name','fc1')  
    concatenationLayer(3,2,'Name','cat1')
```

```

fullyConnectedLayer(hiddenLayerSize1, 'Name', 'fc2')
reluLayer('Name', 'relu2')
fullyConnectedLayer(hiddenLayerSize2, 'Name', 'fc3')
reluLayer('Name', 'relu3')
fullyConnectedLayer(1, 'Name', 'fc4')
tanhLayer('Name', 'tanh1')
scalingLayer('Name', 'scale1', 'Scale', max(actInfo.UpperLimit))
];
dthetaPath = [
    imageInputLayer(obsInfo(2).Dimension, 'Normalization', 'none', 'Name', obsInfo(2).Name)
    fullyConnectedLayer(1, 'Name', 'fc5', 'BiasLearnRateFactor', 0, 'Bias', 0)
];

```

actorNetwork = layerGraph(imgPath);
actorNetwork = addLayers(actorNetwork,dthetaPath);
actorNetwork = connectLayers(actorNetwork,'fc5','cat1/in2');

actorOptions = rlRepresentationOptions('LearnRate',1e-04,'GradientThreshold',1);

Uncomment the following line to use the GPU to accelerate training of the actor CNN.

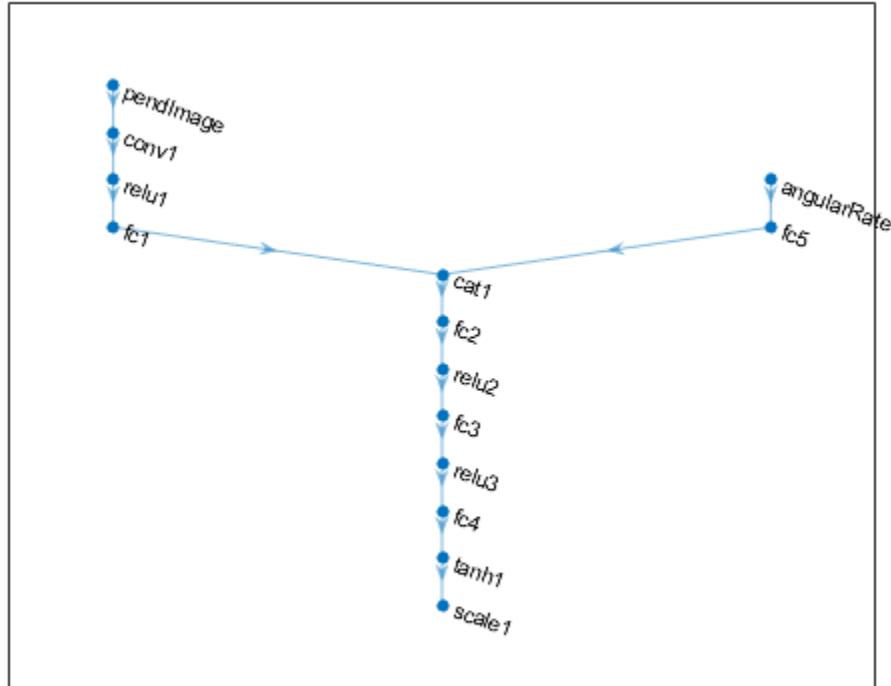
```
% actorOptions.UseDevice = 'gpu';
```

Create the actor representation using the specified neural network and options. For more information, see [rlDeterministicActorRepresentation](#).

```
actor = rlDeterministicActorRepresentation(actorNetwork,obsInfo,actInfo,'Observation',{ 'pendImage'
```

View the actor network configuration.

```
figure
plot(actorNetwork)
```



To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```

agentOptions = rlDDPGAgentOptions(...
    'SampleTime',env.Ts, ...
    'TargetSmoothFactor',1e-3, ...
    'ExperienceBufferLength',1e6, ...
    'DiscountFactor',0.99, ...
    'MiniBatchSize',128);
agentOptions.NoiseOptions.Variance = 0.6;
agentOptions.NoiseOptions.VarianceDecayRate = 1e-6;
  
```

Then, create the agent using the specified actor representation, critic representation, and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

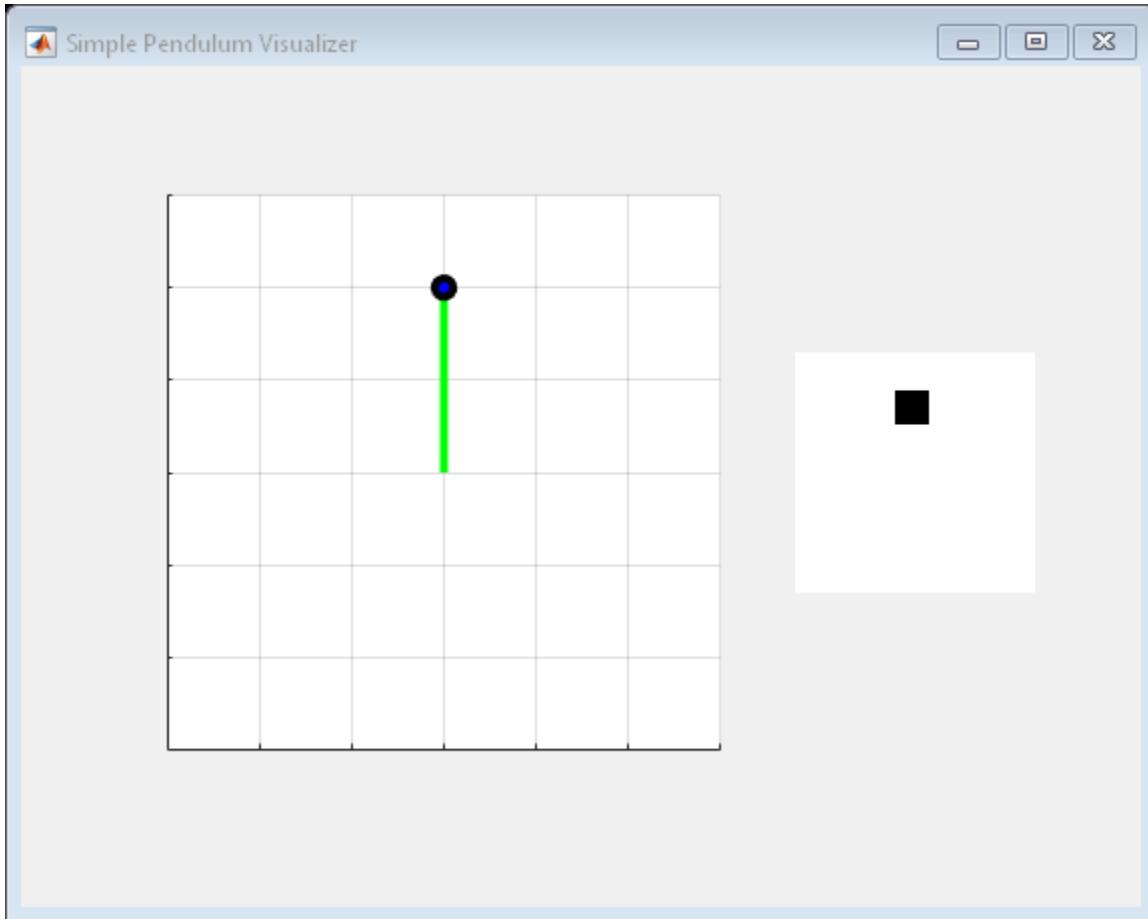
- Run each training for at most 5000 episodes, with each episode lasting at most 400 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option).
- Stop training when the agent receives a moving average cumulative reward greater than -740 (over ten consecutive episodes)

For more information, see `rlTrainingOptions`.

```
maxepisodes = 5000;
maxsteps = 400;
trainingOptions = rlTrainingOptions(...);
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','AverageReward', ...
    'StopTrainingValue',-740);
```

The pendulum system can be visualized with `plot` during training or simulation.

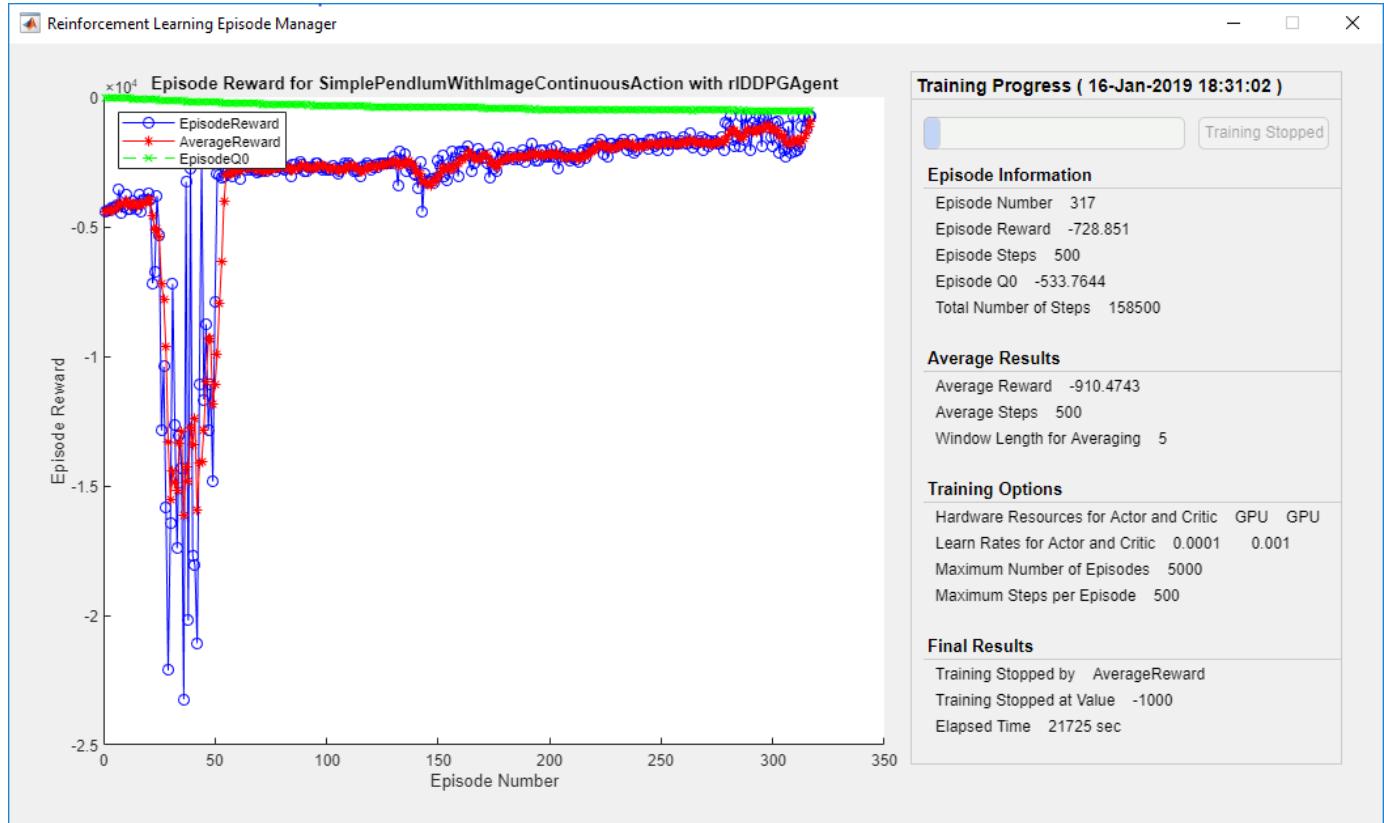
```
plot(env)
```



Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOptions);
else
    % Load pretrained agent for the example.
    load('SimplePendulumWithImageDDPG.mat','agent')
end
```

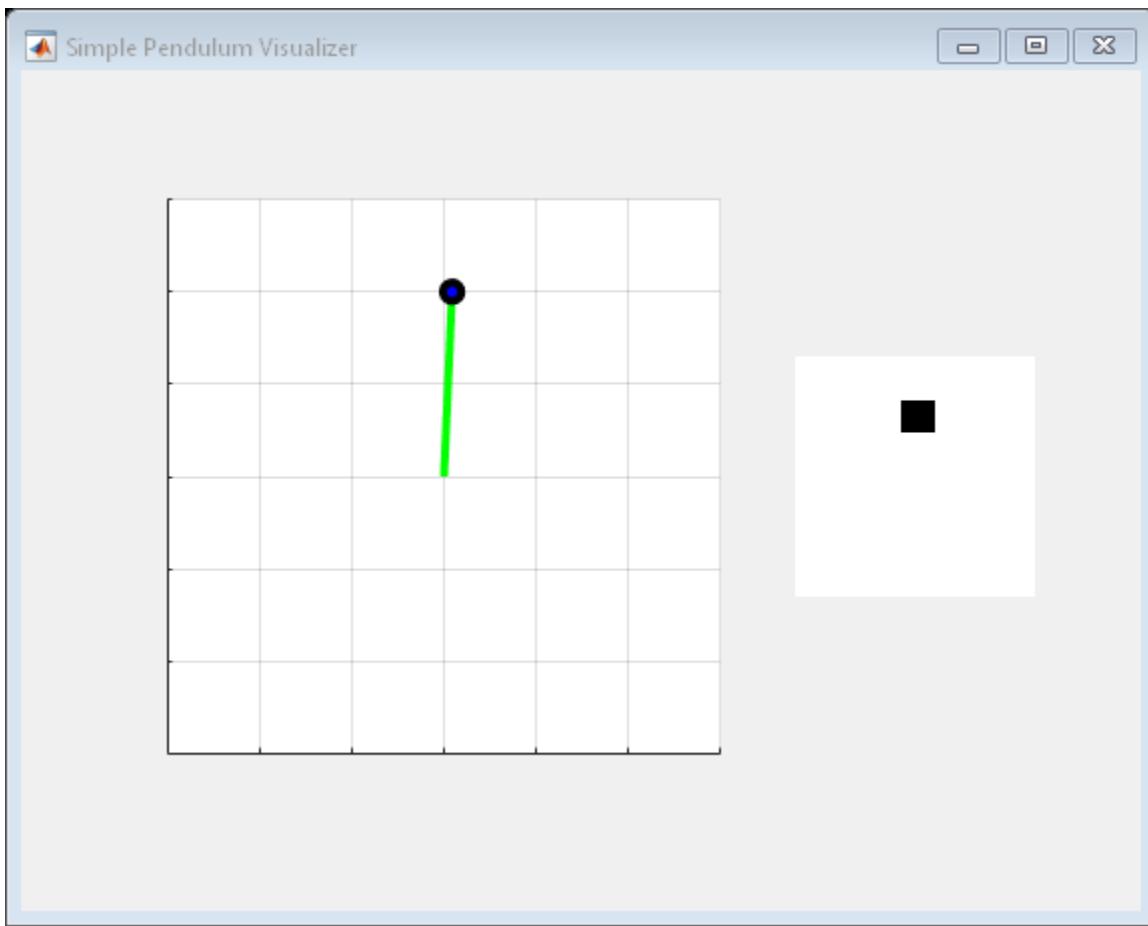
5 Train and Validate Agents



Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



See Also

`train`

More About

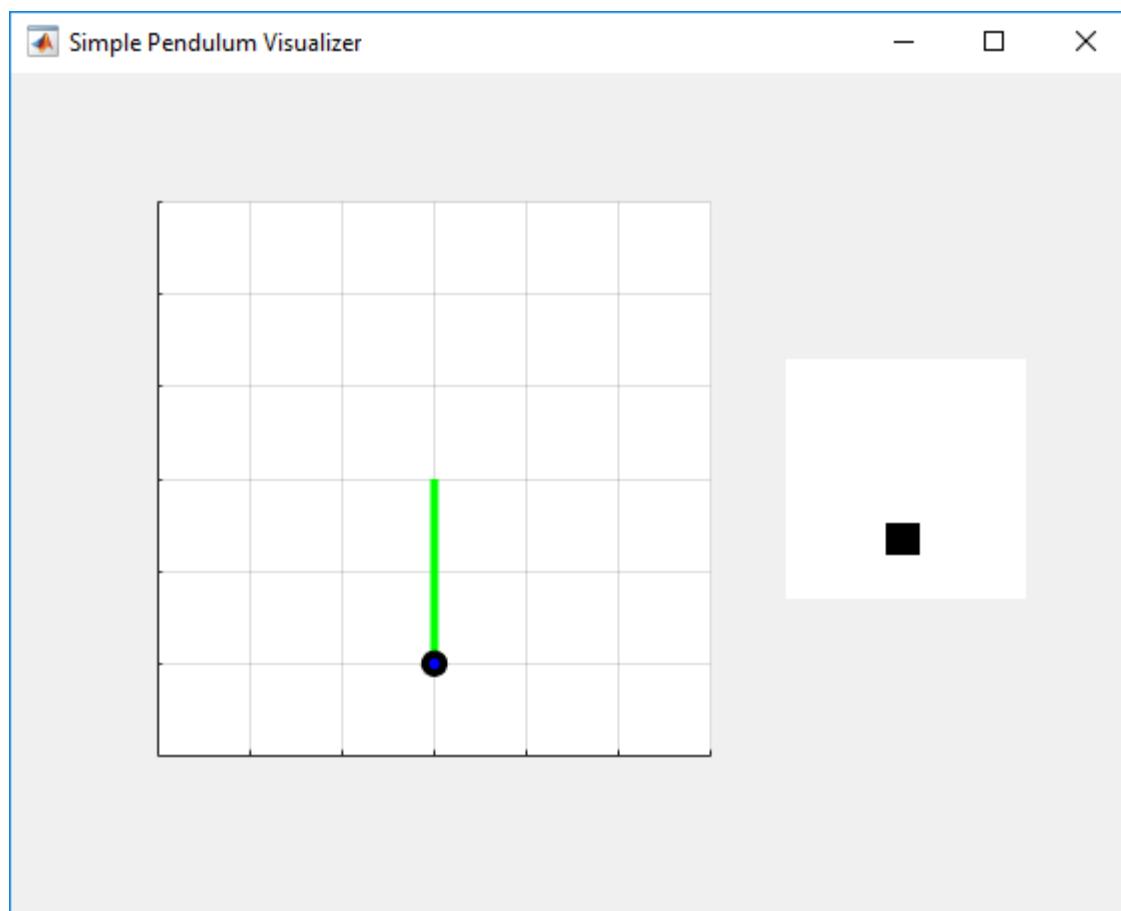
- “Deep Deterministic Policy Gradient Agents” on page 4-14
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Create Agent Using Deep Network Designer and Train Using Image Observations

This example shows how to create a deep Q-learning network (DQN) agent using the Deep Network Designer app to swing up and balance a pendulum modeled in MATLAB®. For more information on DQN agents, see “Deep Q-Network Agents” on page 4-8.

Pendulum Swing Up with image MATLAB Environment

The reinforcement learning environment for this example is a simple frictionless pendulum that is initially hanging in a downward position. The training goal is to make the pendulum stand upright without falling over using minimal control effort.



For this environment:

- The upward balanced pendulum position is 0 radians, and the downward hanging position is π radians.
- The torque action signal from the agent to the environment is from -2 to 2 Nm.
- The observations from the environment are the simplified grayscale image of the pendulum and the pendulum angle derivative.
- The reward r_t , provided at every time step, is:

$$r_t = -\left(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001u_{t-1}^2\right)$$

where:

- θ_t is the angle of displacement from the upright position
- $\dot{\theta}_t$ is the derivative of the displacement angle
- u_{t-1} is the control effort from the previous time step.

For more information on this model, see “Train DDPG Agent to Swing Up and Balance Pendulum with Image Observation” on page 5-64.

Create Environment Interface

Create a predefined environment interface for the pendulum.

```
env = rlPredefinedEnv('SimplePendulumWithImage-Discrete');
```

The interface has two observations. The first observation, named "pendImage", is a 50x50 grayscale image.

```
obsInfo = getObservationInfo(env);
obsInfo(1)

ans =
rlNumericSpec with properties:

    LowerLimit: 0
    UpperLimit: 1
        Name: "pendImage"
    Description: [0x0 string]
    Dimension: [50 50]
    DataType: "double"
```

The second observation, named "angularRate", is the angular velocity of the pendulum.

```
obsInfo(2)

ans =
rlNumericSpec with properties:

    LowerLimit: -Inf
    UpperLimit: Inf
        Name: "angularRate"
    Description: [0x0 string]
    Dimension: [1 1]
    DataType: "double"
```

The interface has a discrete action space where the agent can apply one of five possible torque values to the pendulum: -2, -1, 0, 1, or 2 Nm.

```
actInfo = getActionInfo(env)

actInfo =
rlFiniteSetSpec with properties:
```

```
Elements: [-2 -1 0 1 2]
Name: "torque"
Description: [0x0 string]
Dimension: [1 1]
DataType: "double"
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

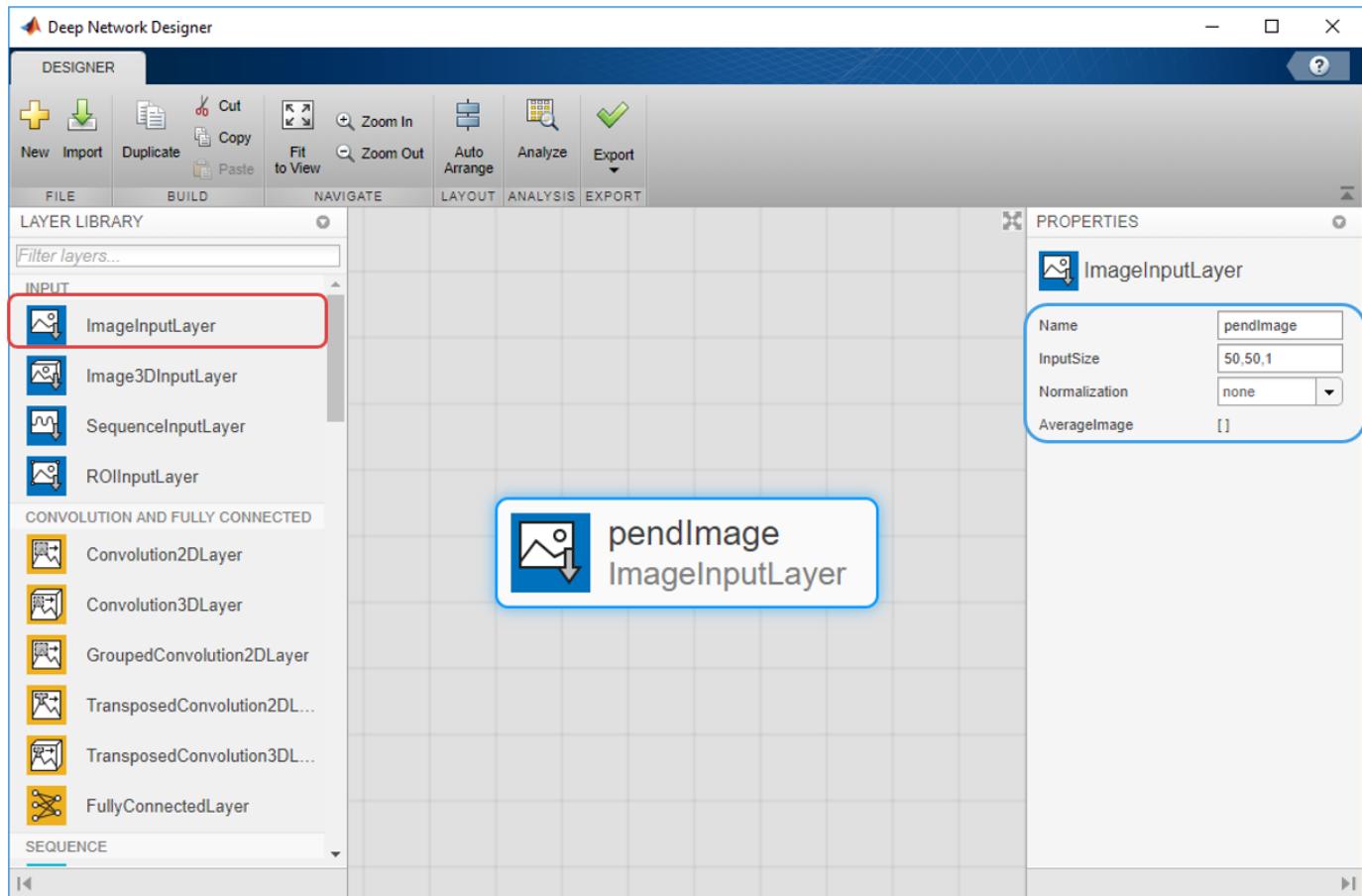
Construct Critic Network Using Deep Network Designer

A DQN agent approximates the long-term reward given observations and actions using a critic value function representation. For this environment, the critic is a deep neural network with three inputs, the two observations and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

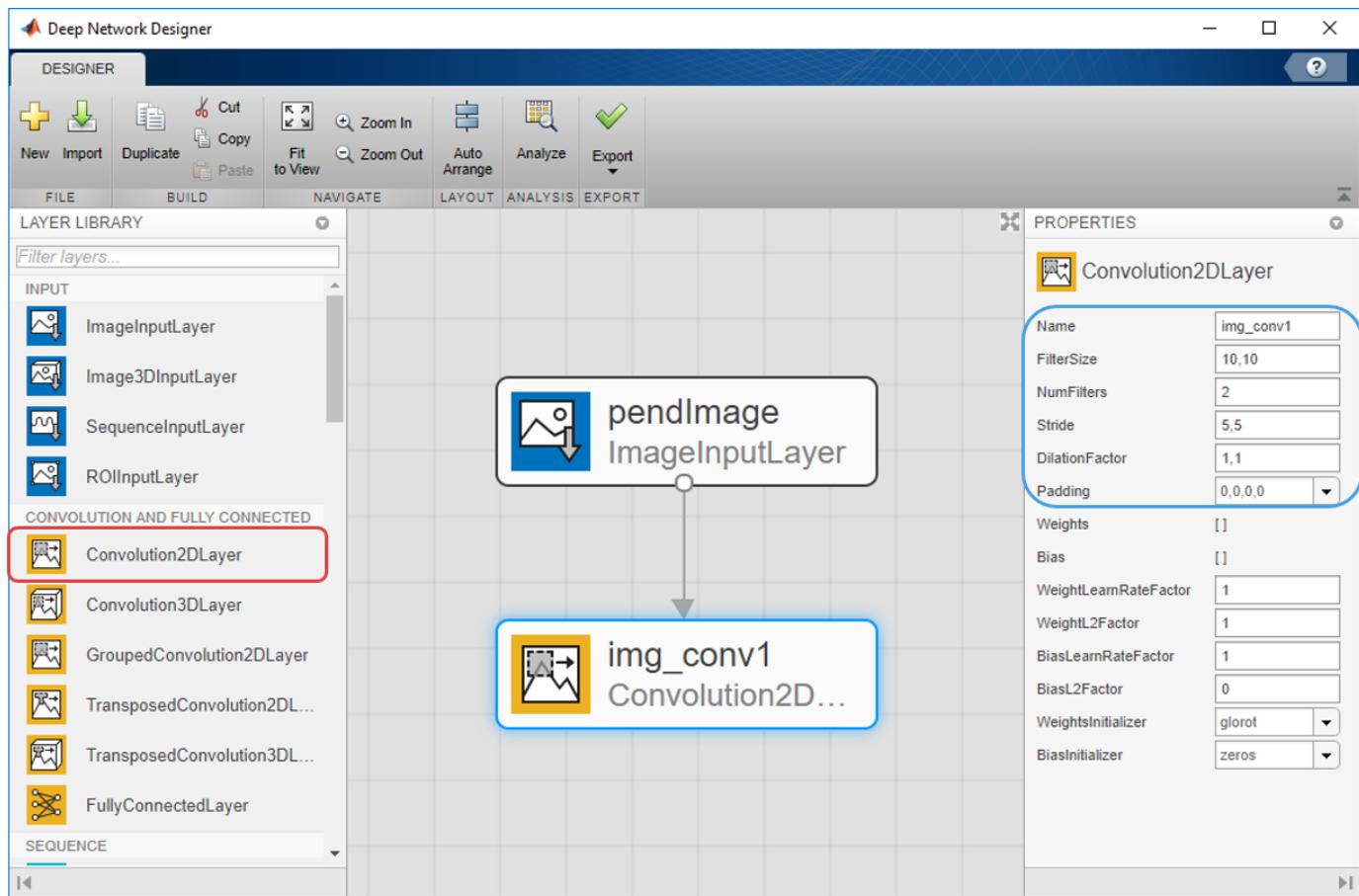
You can interactively construct the critic network using the Deep Network Designer app. To do so, you first create separate input paths for each observation and action. These paths learn lower-level features from their respective inputs. You then create a common output path which combines the outputs from the input paths.

Create Image Observation Path

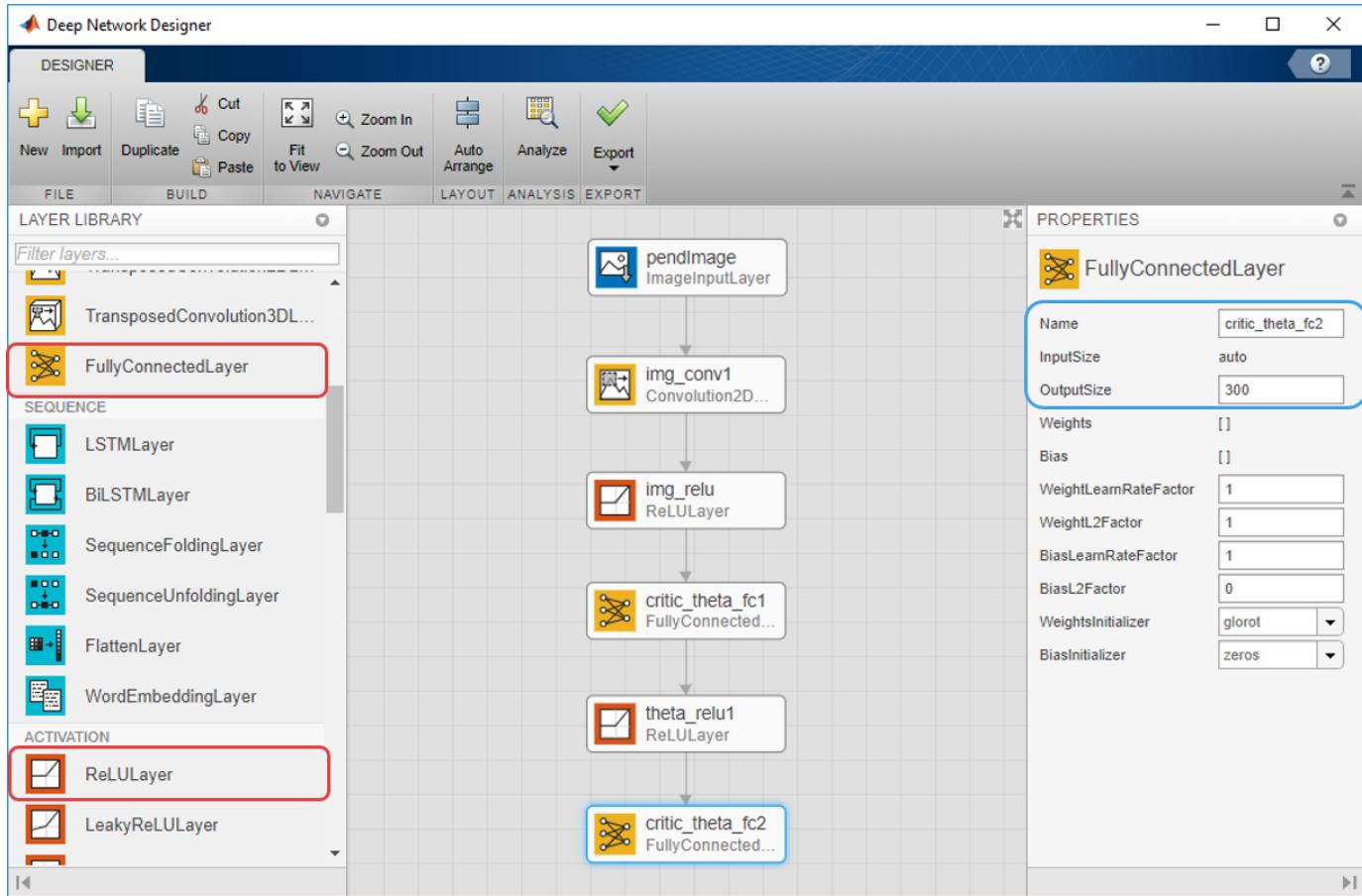
To create the image observation path, first drag an **ImageInputLayer** from the **Layer Library** pane to the canvas. Set the layer **InputSize** to **50,50,1** for the image observation, and set **Normalization** to **none**.



Second, drag a Convolution2DLayer to the canvas and connect the input of this layer to the output of the ImageInputLayer. Create a convolution layer with 2 filters (**NumFilters** property) that have a height and width of 10 (**FilterSize** property), and use a stride of 5 in the horizontal and vertical directions (**Stride** property).



Finally, complete the image path network with two sets of ReLU¹Layer and FullyConnectedLayer. The **OutputSize** of the two FullyConnectedLayer layers are 400 and 300, respectively.



Create All Input Paths and Output Path

Construct the other input paths and output path in similar fashion. For this example, use the following options:

Angular Velocity Path (scalar input):

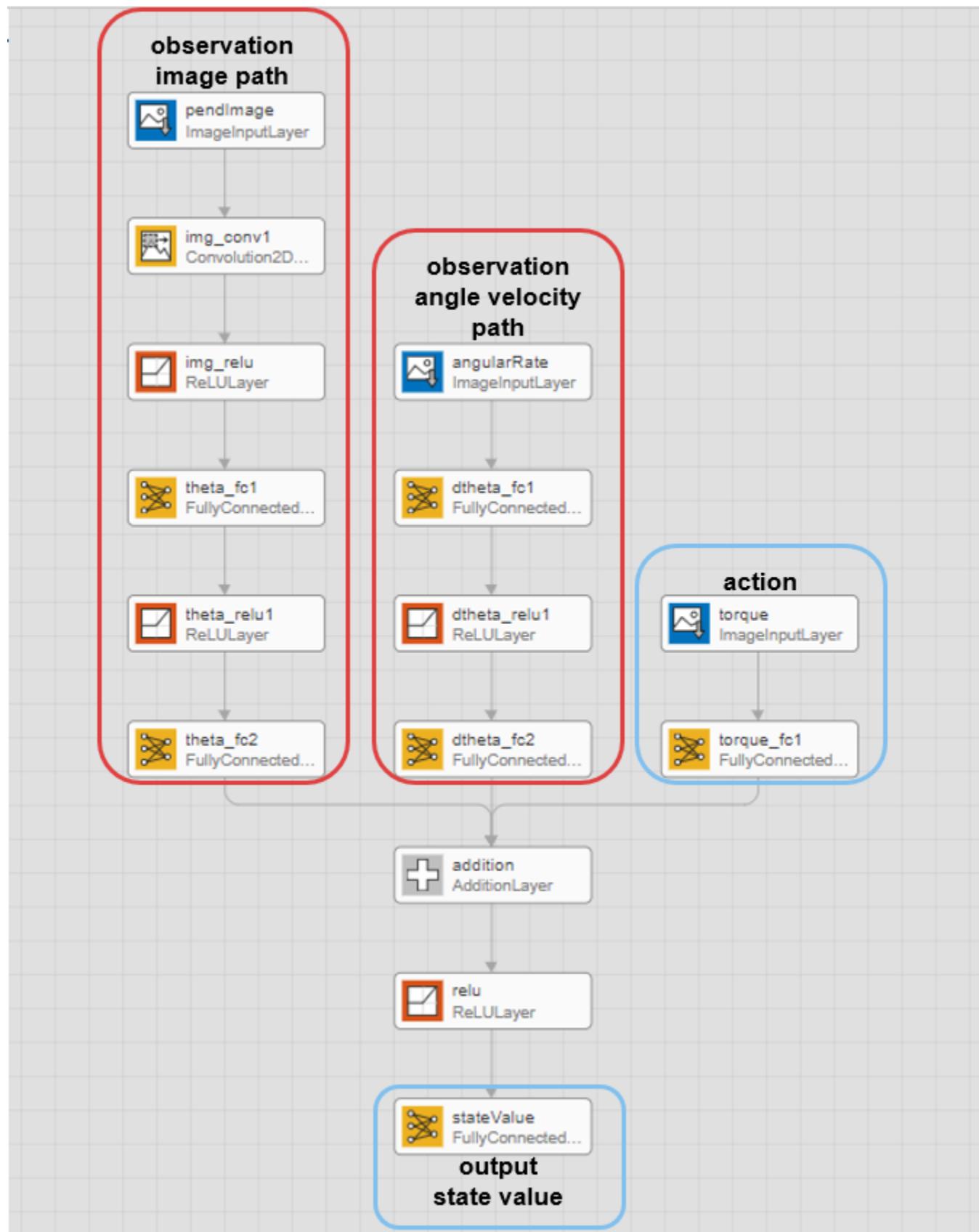
- **ImageInputLayer: InputSize = 1,1** and **Normalization = none**
- **FullyConnectedLayer: OutputSize = 400**
- **ReLUlayer**
- **FullyConnectedLayer: OutputSize = 300**

Action Path (scalar input):

- **ImageInputLayer: InputSize = 1,1** and **Normalization = none**
- **FullyConnectedLayer: OutputSize = 300**

Output Path:

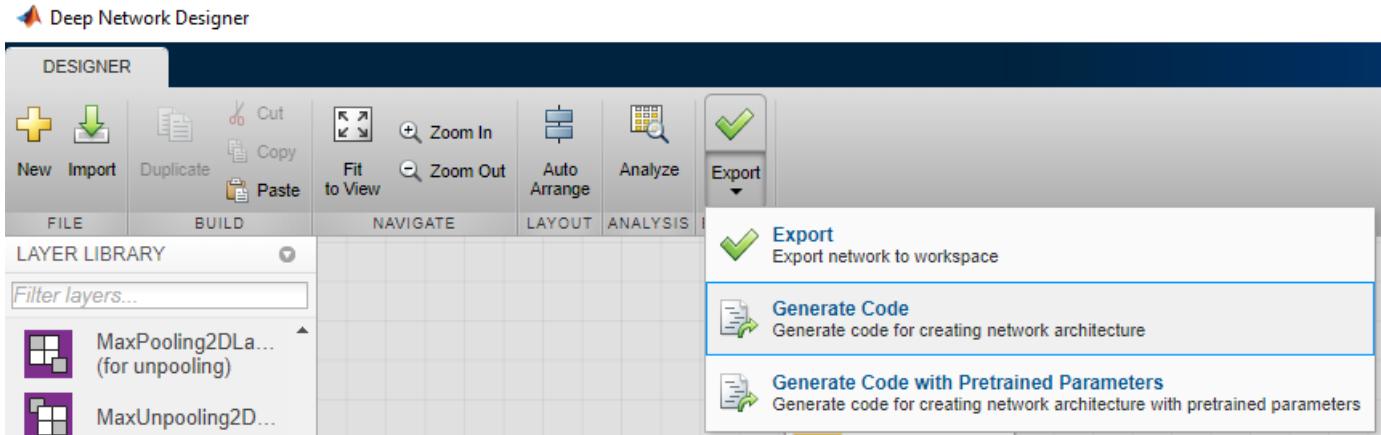
- **AdditionLayer**: Connect the output of all input paths to the input of this layer.
- **ReLUlayer**
- **FullyConnectedLayer: OutputSize = 1** for the scalar value function.



Export Network from Deep Network Designer

To export the network to the MATLAB workspace, in the **Deep Network Designer**, click **Export**. The **Deep Network Designer** exports the network to a new variable containing the network layers. You can create the critic representation using this layer network variable.

Alternatively, to generate equivalent MATLAB code for the network, click **Export > Generate Code**.

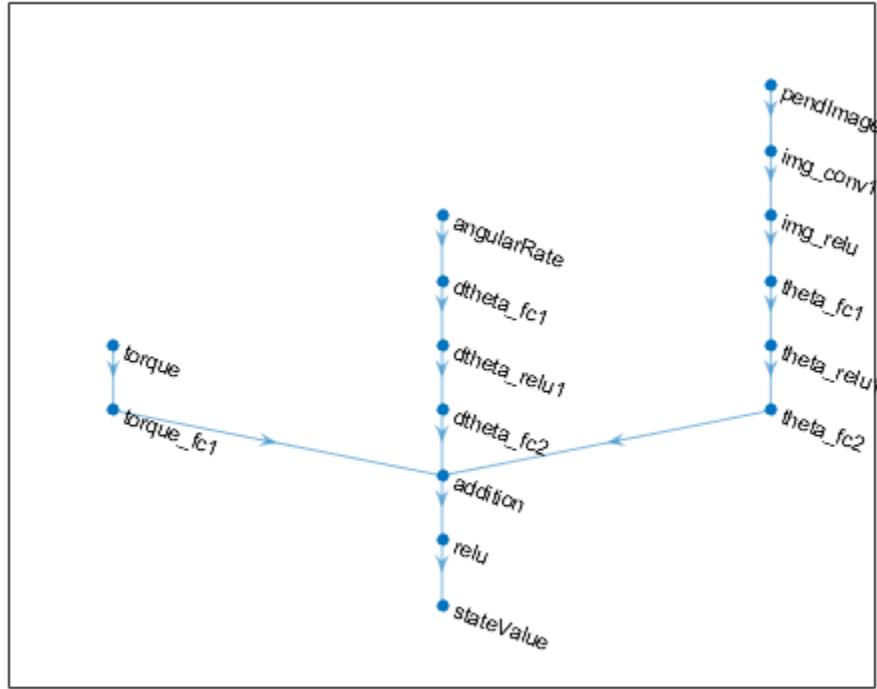


The generated code is:

```
lgraph = layerGraph();
layers = [
    imageInputLayer([1 1 1], "Name", "torque", "Normalization", "none")
    fullyConnectedLayer(300, "Name", "torque_fc1"]);
lgraph = addLayers(lgraph, layers);
layers = [
    imageInputLayer([1 1 1], "Name", "angularRate", "Normalization", "none")
    fullyConnectedLayer(400, "Name", "dtheta_fc1")
    reluLayer("Name", "dtheta_relu1")
    fullyConnectedLayer(300, "Name", "dtheta_fc2"));
lgraph = addLayers(lgraph, layers);
layers = [
    imageInputLayer([50 50 1], "Name", "pendImage", "Normalization", "none")
    convolution2dLayer([10 10], 2, "Name", "img_conv1", "Stride", [5 5])
    reluLayer("Name", "img_relu")
    fullyConnectedLayer(400, "Name", "theta_fc1")
    reluLayer("Name", "theta_relu1")
    fullyConnectedLayer(300, "Name", "theta_fc2"));
lgraph = addLayers(lgraph, layers);
layers = [
    additionLayer(3, "Name", "addition")
    reluLayer("Name", "relu")
    fullyConnectedLayer(1, "Name", "stateValue"));
lgraph = addLayers(lgraph, layers);
lgraph = connectLayers(lgraph, "torque_fc1", "addition/in3");
lgraph = connectLayers(lgraph, "theta_fc2", "addition/in1");
lgraph = connectLayers(lgraph, "dtheta_fc2", "addition/in2");
```

View the critic network configuration.

```
figure
plot(lgraph)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate',1e-03,'GradientThreshold',1);
```

Create the critic representation using the specified deep neural network `lgraph` and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(lgraph,obsInfo,actInfo,...  
    'Observation',{'pendImage','angularRate'},'Action',{'torque'},criticOpts);
```

To create the DQN agent, first specify the DQN agent options using `rldQNAgentOptions`.

```
agentOpts = rldQNAgentOptions(...  
    'UseDoubleDQN',false,...  
    'TargetUpdateMethod',"smoothing",...  
    'TargetSmoothFactor',1e-3,...  
    'ExperienceBufferLength',1e6,...  
    'DiscountFactor',0.99,...  
    'SampleTime',env.Ts,...  
    'MiniBatchSize',64);  
agentOpts.EpsilonGreedyExploration.EpsilonDecay = 1e-5;
```

Then, create the DQN agent using the specified critic representation and agent options. For more information, see `rldQNAgent`.

```
agent = rldQNAgent(critic,agentOpts);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

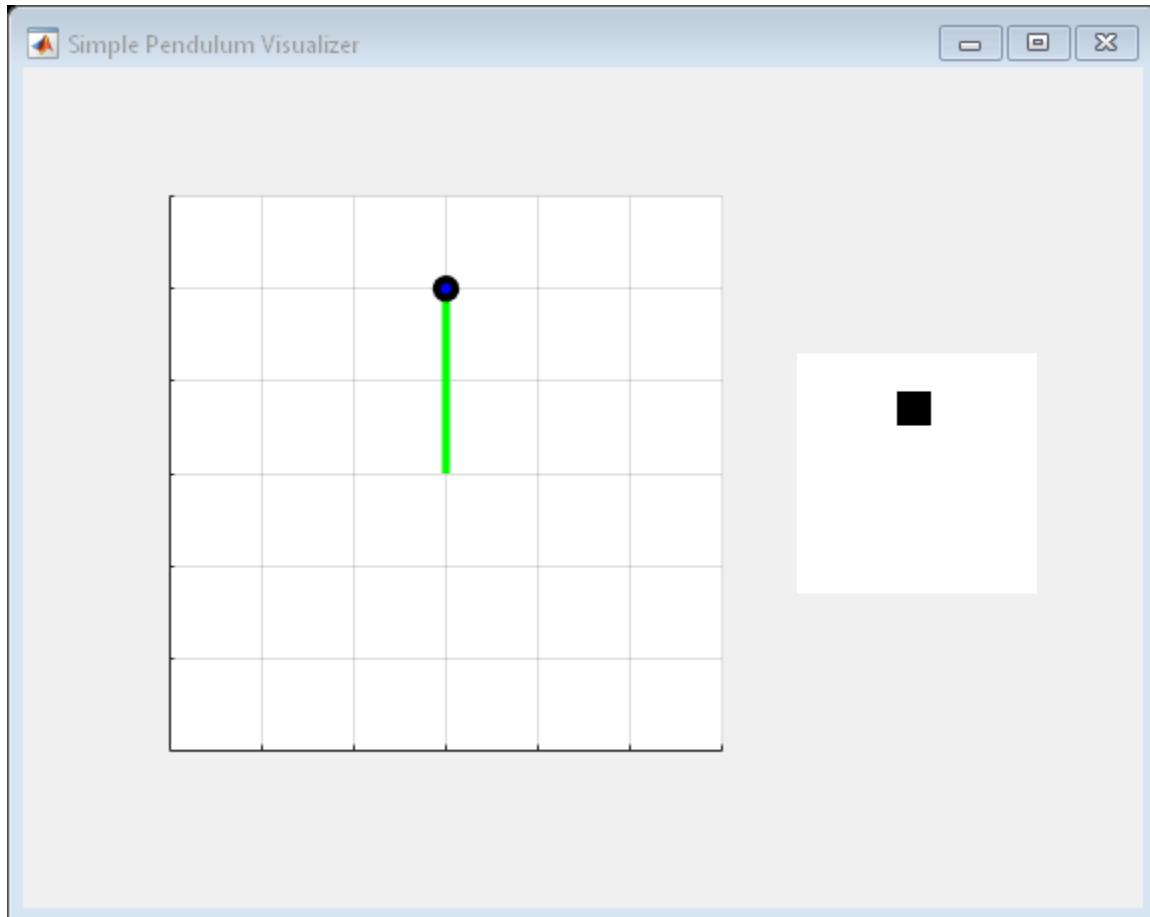
- Run each training for at most 1000 episodes, with each episode lasting at most 500 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than `-1000` over five consecutive episodes. At this point, the agent can quickly balance the pendulum in the upright position using minimal control effort.

For more information, see `rlTrainingOptions`.

```
trainOpts = rlTrainingOptions(...  
    'MaxEpisodes',5000,...  
    'MaxStepsPerEpisode',500,...  
    'Verbose',false,...  
    'Plots','training-progress',...  
    'StopTrainingCriteria','AverageReward',...  
    'StopTrainingValue',-1000);
```

The pendulum system can be visualized with `plot(env)` during training or simulation.

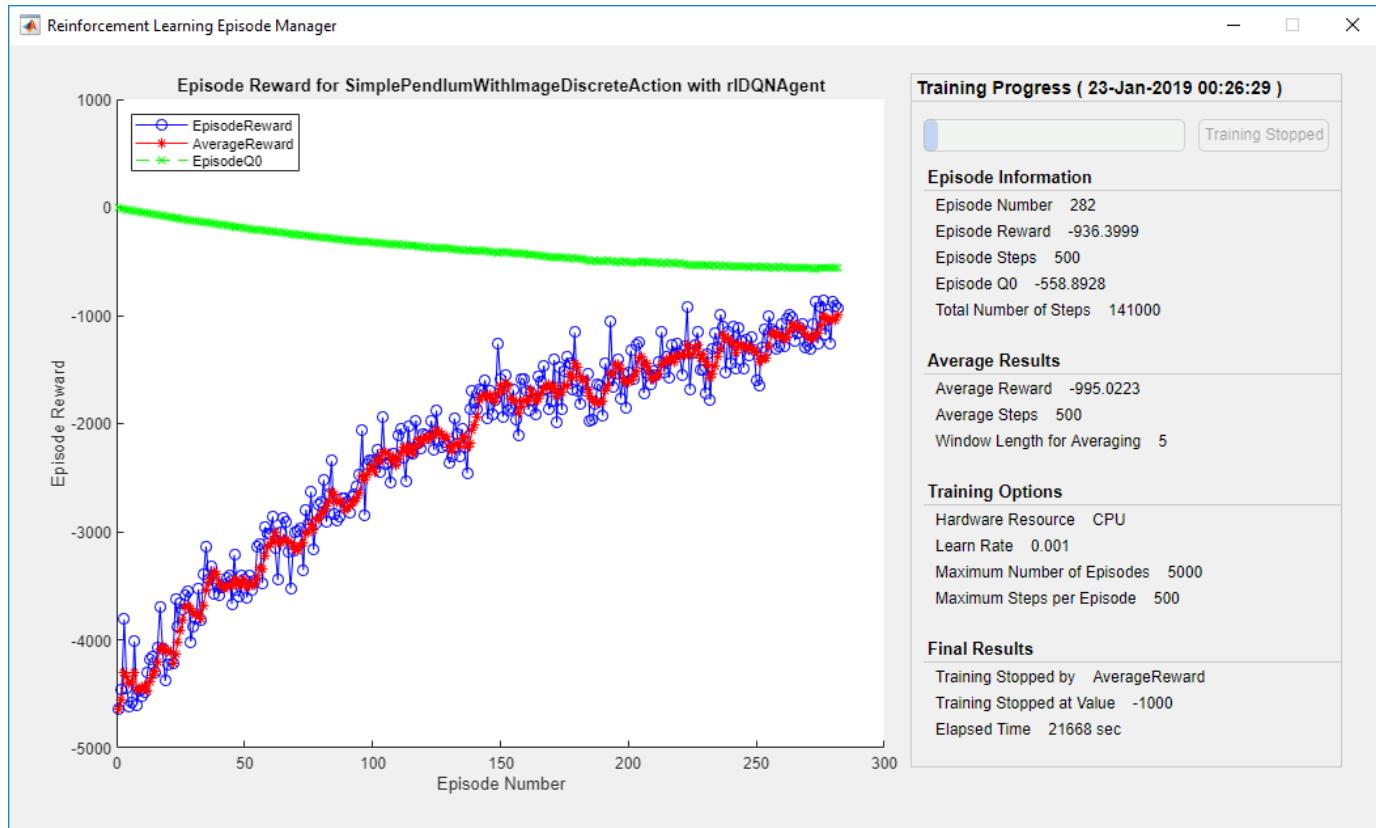
```
plot(env)
```



5 Train and Validate Agents

Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

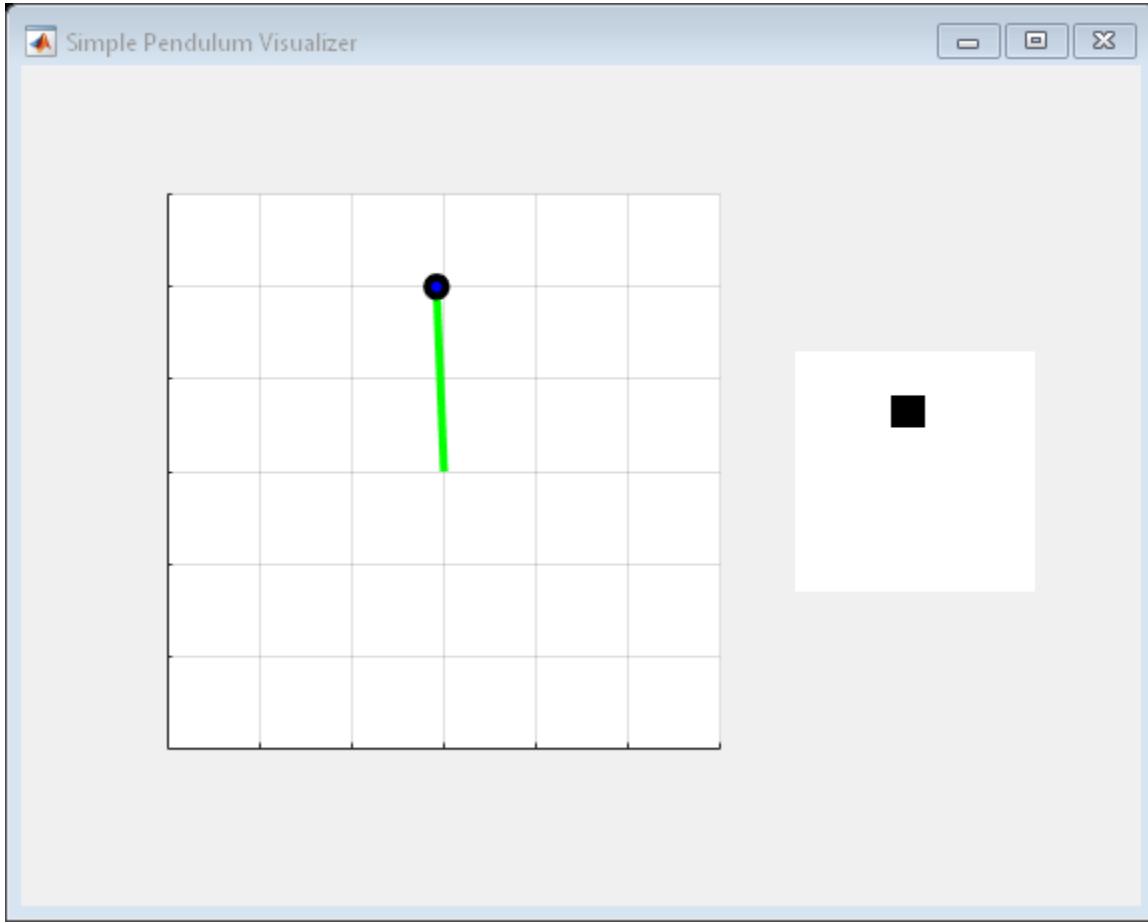
```
doTraining = false;  
  
if doTraining  
    % Train the agent.  
    trainingStats = train(agent,env,trainOpts);  
else  
    % Load pretrained agent for the example.  
    load('MATLABPendImageDQN.mat','agent');  
end
```



Simulate DQN Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);  
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)  
totalReward = -888.9802
```

See Also

[Deep Network Designer](#) | [rlDQNAgent](#)

More About

- “Train DQN Agent to Swing Up and Balance Pendulum” on page 5-37

Train AC Agent to Balance Cart-Pole System Using Parallel Computing

This example shows how to train an actor-critic (AC) agent to balance a cart-pole system modeled in MATLAB® by using asynchronous parallel training. For an example that shows how to train the agent without using parallel training, see “Train AC Agent to Balance Cart-Pole System” on page 5-19.

Actor Parallel Training

When you use parallel computing with AC agents, each worker generates experiences from its copy of the agent and the environment. After every N steps, the worker computes gradients from the experiences and sends the computed gradients back to the host agent. The host agent updates its parameters as follows.

- For asynchronous training, the host agent applies the received gradients without waiting for all workers to send gradients, and sends the updated parameters back to the worker that provided the gradients. Then, the worker continues to generate experiences from its environment using the updated parameters.
- For synchronous training, the host agent waits to receive gradients from all of the workers and updates its parameters using these gradients. The host then sends updated parameters to all the workers at the same time. Then, all workers continue to generate experiences using the updated parameters.

Create Cart-Pole MATLAB Environment Interface

Create a predefined environment interface for the cart-pole system. For more information on this environment, see “Load Predefined Control System Environments” on page 2-15.

```
env = rlPredefinedEnv("CartPole-Discrete");
env.PenaltyForFalling = -10;
```

Obtain the observation and action information from the environment interface.

```
obsInfo = getObservationInfo(env);
numObservations = obsInfo.Dimension(1);
actInfo = getActionInfo(env);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create AC Agent

An AC agent approximates the long-term reward, given observations and actions, using a critic value function representation. To create the critic, first create a deep neural network with one input (the observation) and one output (the state value). The input size of the critic network is [4 1 1] since the environment provides 4 observations. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
criticNetwork = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(32, 'Name', 'CriticStateFC1')
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(1, 'Name', 'CriticFC')];
```

```

criticOpts = rlRepresentationOptions('LearnRate',1e-2,'GradientThreshold',1);
critic = rlValueRepresentation(criticNetwork,obsInfo,'Observation',{'state'},criticOpts);

```

An AC agent decides which action to take, given observations, using an actor representation. To create the actor, create a deep neural network with one input (the observation) and one output (the action). The output size of the actor network is 2 since the agent can apply 2 force values to the environment, -10 and 10.

```

actorNetwork = [
    imageInputLayer([4 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(32, 'Name', 'ActorStateFC1')
    reluLayer('Name', 'ActorRelu1')
    fullyConnectedLayer(2, 'Name', 'action')];

actorOpts = rlRepresentationOptions('LearnRate',1e-2,'GradientThreshold',1);

actor = rlStochasticActorRepresentation(actorNetwork,obsInfo,actInfo,....
    'Observation',{'state'},actorOpts);

```

To create the AC agent, first specify the AC agent options using `rlACAgentOptions`.

```

agentOpts = rlACAgentOptions(...
    'NumStepsToLookAhead',32,...
    'EntropyLossWeight',0.01,...
    'DiscountFactor',0.99);

```

Then create the agent using the specified actor representation and agent options. For more information, see `rlACAgent`.

```
agent = rlACAgent(actor,critic,agentOpts);
```

Parallel Training Options

To train the agent, first specify the training options. For this example, use the following options.

- Run each training for at most 1000 episodes, with each episode lasting at most 500 time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option).
- Stop training when the agent receives an average cumulative reward greater than 500 over 10 consecutive episodes. At this point, the agent can balance the pendulum in the upright position.

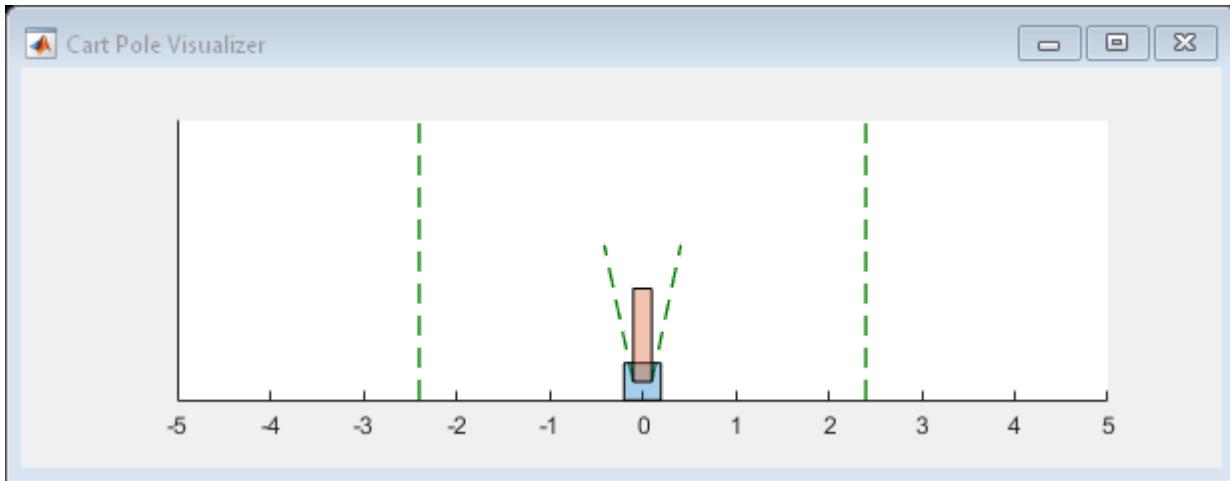
```

trainOpts = rlTrainingOptions(...
    'MaxEpisodes',1000,....
    'MaxStepsPerEpisode', 500,....
    'Verbose',false,....
    'Plots','training-progress',...
    'StopTrainingCriteria','AverageReward',...
    'StopTrainingValue',500,....
    'ScoreAveragingWindowLength',10);

```

You can visualize the cart-pole system during training or simulation using the `plot` function.

```
plot(env)
```



To train the agent using parallel computing, specify the following training options.

- Set the `UseParallel` option to `True`.
- Train the agent in parallel asynchronously by setting the `ParallelizationOptions.Mode` option to `"async"`.
- After every 32 steps, each worker computes gradients from experiences and send them to the host.
- The AC agent requires workers to send `"gradients"` to the host.
- The AC agent requires `'StepsUntilDataIsSent'` to be equal to `agentOptions.NumStepsToLookAhead`.

```
trainOpts.UseParallel = true;
trainOpts.ParallelizationOptions.Mode = "async";
trainOpts.ParallelizationOptions.DataToSendFromWorkers = "gradients";
trainOpts.ParallelizationOptions.StepsUntilDataIsSent = 32;
```

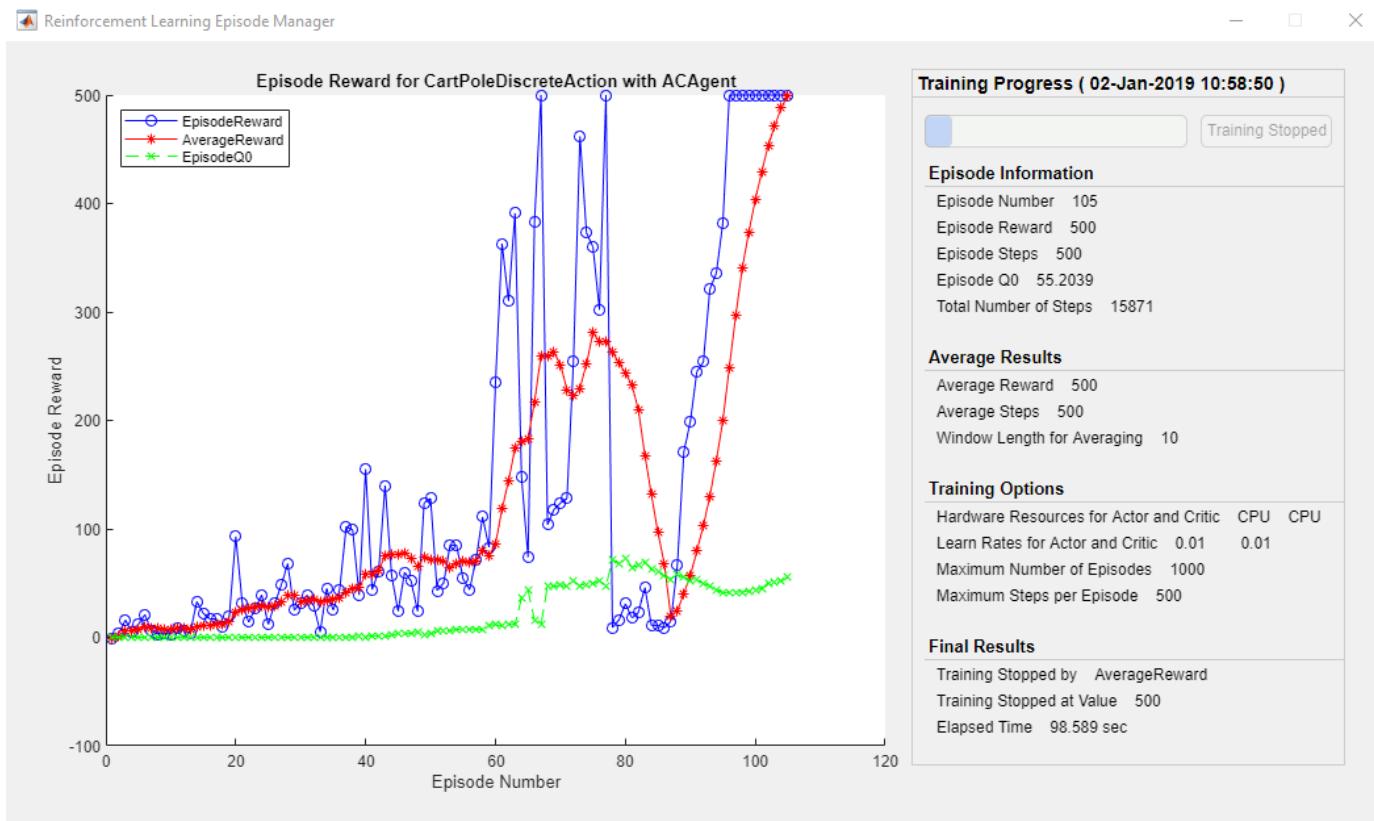
For more information, see `rlTrainingOptions`.

Train Agent

Train the agent using the `train` function. Training the agent is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`. Due to randomness in the asynchronous parallel training, you can expect different training results from the following training plot. The plot shows the result of training with six workers.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load the pretrained agent for the example.
    load('MATLABCartpoleParAC.mat','agent');
end
```



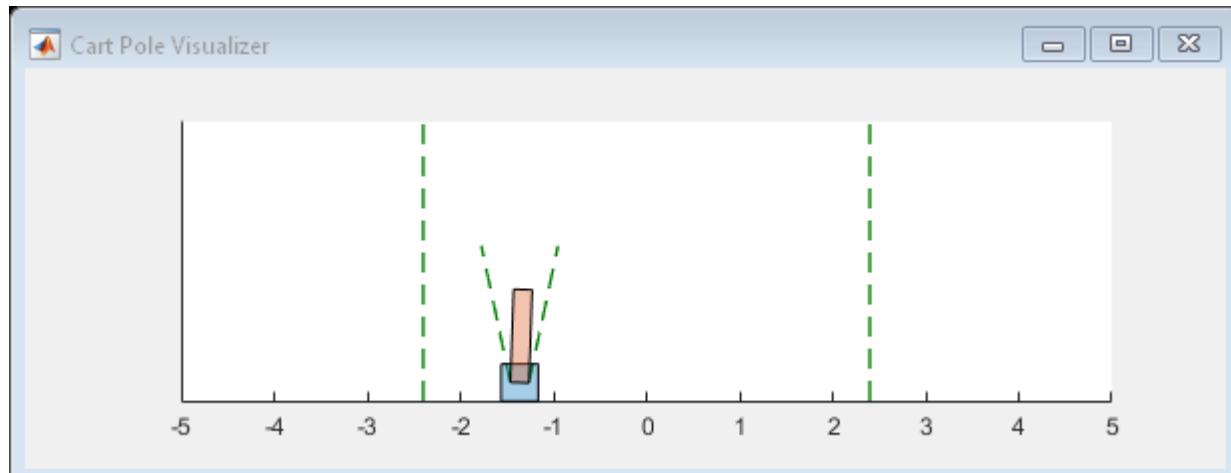
Simulate AC Agent

You can visualize the cart-pole system with the `plot` function during simulation.

```
plot(env)
```

To validate the performance of the trained agent, simulate it within the cart-pole environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',500);
experience = sim(env,agent,simOptions);
```



```
totalReward = sum(experience.Reward)  
totalReward = 500
```

References

[1] Mnih, Volodymyr, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 'Asynchronous Methods for Deep Reinforcement Learning'. *ArXiv:1602.01783 [Cs]*, 16 June 2016. <https://arxiv.org/abs/1602.01783>.

See Also

[rlTrainingOptions](#) | [train](#)

More About

- "Train Reinforcement Learning Agents" on page 5-2

Train DDPG Agent to Control Flying Robot

This example shows how to train a deep deterministic policy gradient (DDPG) agent to generate trajectories for a flying robot modeled in Simulink®. For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14.

Flying Robot Model

The reinforcement learning environment for this example is a flying robot with its initial condition randomized around a ring of radius 15 m where the orientation of the robot is also randomized. The robot has two thrusters mounted on the side of the body which are used to propel and steer the robot. The training goal is to drive the robot from its initial condition to the origin facing east.

Open the model and setup initial model variables.

```
mdl = 'rlFlyingRobotEnv';
open_system(mdl)

% initial model state variables
theta0 = 0;
x0 = -15;
y0 = 0;

% sample time
Ts = 0.4;

% simulation length
Tf = 30;
```

For this model:

- The goal orientation is θ radians (robot facing east).
- The thrust from each actuator is bounded from -1 to 1 N
- The observations from the environment are the position, orientation (sine and cosine of orientation), velocity and angular velocity of the robot.
- The reward r_t , provided at every time step is:

$$r_1 = 10((x_t^2 + y_t^2 + \theta_t^2) < 0.5)$$

$$r_2 = -100(|x_t| \geq 20 \text{ || } |y_t| \geq 20)$$

$$r_3 = -\left(0.2(R_{t-1} + L_{t-1})^2 + 0.3(R_{t-1} - L_{t-1})^2 + 0.03x_t^2 + 0.03y_t^2 + 0.02\theta_t^2\right)$$

$$r_t = r_1 + r_2 + r_3$$

where:

- x_t is the position of the robot along the x-axis.
- y_t is the position of the robot along the y-axis.
- θ_t is the orientation of the robot.
- L_{t-1} is the control effort from the left thruster.

- R_{t-1} is the control effort from the right thruster.
- r_1 is the reward when the robot is close to the goal.
- r_2 is the penalty when the robot drives beyond 20 m in either the x or y direction. The simulation is terminated when $r_2 < 0$.
- r_3 is a QR penalty that penalizes distance from the goal and control effort.

Create Integrated Model

To train an agent for the `FlyingRobotEnv` model, use the `createIntegratedEnv` function to automatically generate an integrated model with the RL Agent block that is ready for training.

```
integratedMdl = 'IntegratedFlyingRobot';
[~,agentBlk,observationInfo,actionInfo] = createIntegratedEnv(mdl,integratedMdl);
```

Actions and Observations

Before creating the environment object, specify names for the observation and action specifications, and bound the thrust actions between -1 and 1.

The observation signals for this environment are $\text{observation} = [x \ y \ \dot{x} \ \dot{y} \ \sin(\theta) \ \cos(\theta) \ \dot{\theta}]^T$.

```
numObs = prod(observationInfo.Dimension);
observationInfo.Name = 'observations';
```

The action signals for this environment are $\text{action} = [T_R \ T_L]^T$.

```
numAct = prod(actionInfo.Dimension);
actionInfo.LowerLimit = -ones(numAct,1);
actionInfo.UpperLimit = ones(numAct,1);
actionInfo.Name = 'thrusts';
```

Create Environment Interface

Create an environment interface for the flying robot with `rlSimulinkEnv` with the generated model.

```
env = rlSimulinkEnv(integratedMdl,agentBlk,observationInfo,actionInfo);
```

Reset Function

Create a custom reset function that randomizes the initial position of the robot along a ring of radius 15 m and the initial orientation. See `flyingRobotResetFcn` for details of the reset function.

```
env.ResetFcn = @(in) flyingRobotResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs (the observation and action) and one output. For more information on creating a neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
% specify the number of outputs for the hidden layers.
hiddenLayerSize = 100;

observationPath = [
    imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc2')
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(1, 'Name', 'fc4')];
actionPath = [
    imageInputLayer([numAct 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc5')];

% create the layerGraph
criticNetwork = layerGraph(observationPath);
criticNetwork = addLayers(criticNetwork, actionPath);

% connect actionPath to observationPath
criticNetwork = connectLayers(criticNetwork, 'fc5', 'add/in2');
```

Specify options for the critic using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate', 1e-03, 'GradientThreshold', 1);
```

Create the critic representation using the specified neural network and options. You must also specify the action and observation specification for the critic. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork, observationInfo, actionInfo, ...
    'Observation', {'observation'}, 'Action', {'action'}, criticOptions);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input (the observation) and one output (the action).

Construct the actor in a similar manner to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [
    imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc2')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(numAct, 'Name', 'fc4')
    tanhLayer('Name', 'tanh1')];

actorOptions = rlRepresentationOptions('LearnRate', 1e-04, 'GradientThreshold', 1);

actor = rlDeterministicActorRepresentation(actorNetwork, observationInfo, actionInfo, ...
    'Observation', {'observation'}, 'Action', {'tanh1'}, actorOptions);
```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```
agentOptions = rlDDPGAgentOptions(...
    'SampleTime',Ts, ...
    'TargetSmoothFactor',1e-3, ...
    'ExperienceBufferLength',1e6 , ...
    'DiscountFactor',0.99, ...
    'MiniBatchSize',256);
agentOptions.NoiseOptions.Variance = 1e-1;
agentOptions.NoiseOptions.VarianceDecayRate = 1e-6;
```

Then, create the agent using the specified actor representation, critic representation, and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

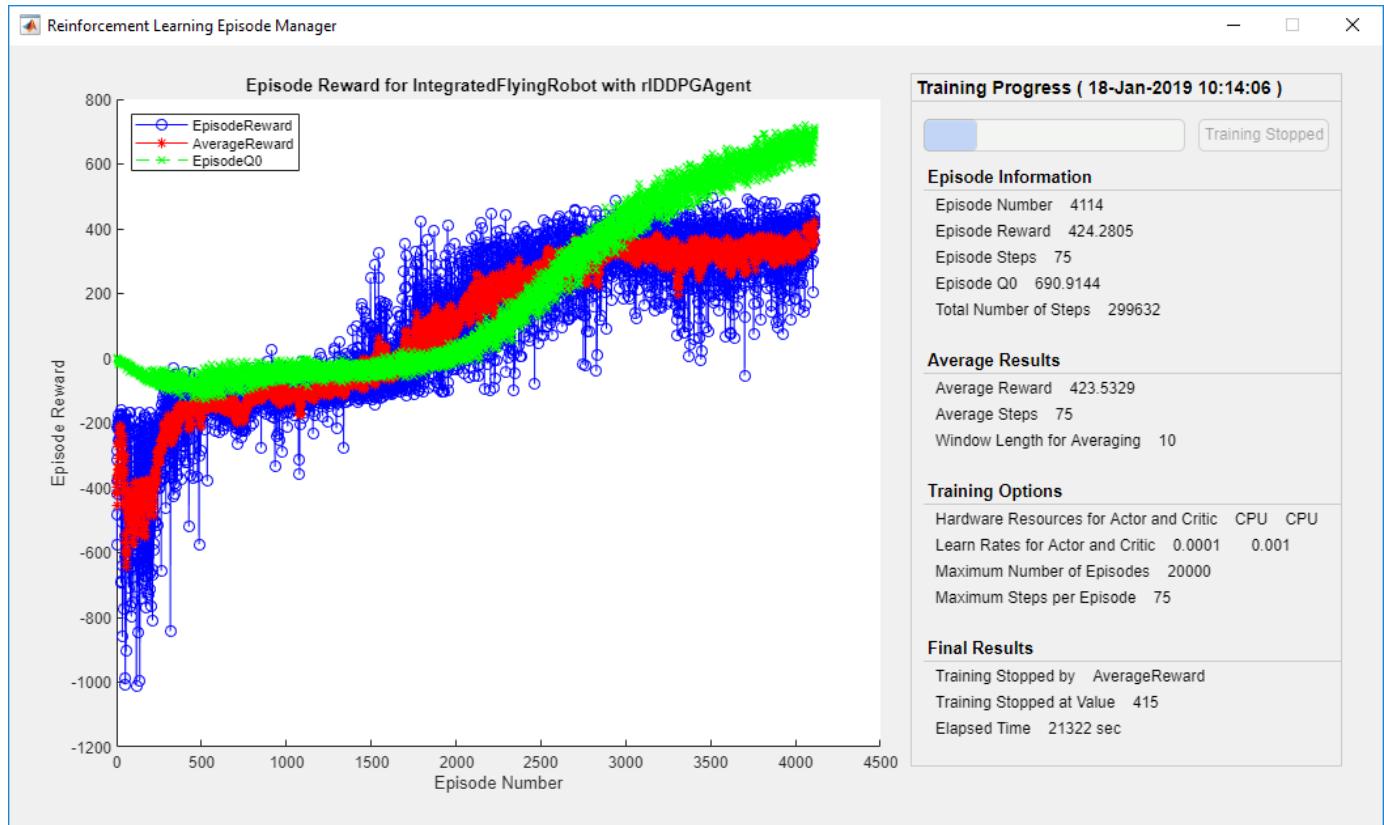
- Run each training for at most 20000 episodes, with each episode lasting at most `ceil(Tf/Ts)` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option to `false`).
- Stop training when the agent receives an average cumulative reward greater than 415 over ten consecutive episodes. At this point, the agent can drive the flying robot to the goal position.
- Save a copy of the agent for each episode where the cumulative reward is greater than 415.

For more information, see `rlTrainingOptions`.

```
maxepisodes = 20000;
maxsteps = ceil(Tf/Ts);
trainingOptions = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'StopOnError','on', ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','AverageReward', ...
    'StopTrainingValue',415, ...
    'ScoreAveragingWindowLength',10, ...
    'SaveAgentCriteria','EpisodeReward', ...
    'SaveAgentValue',415);
```

Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

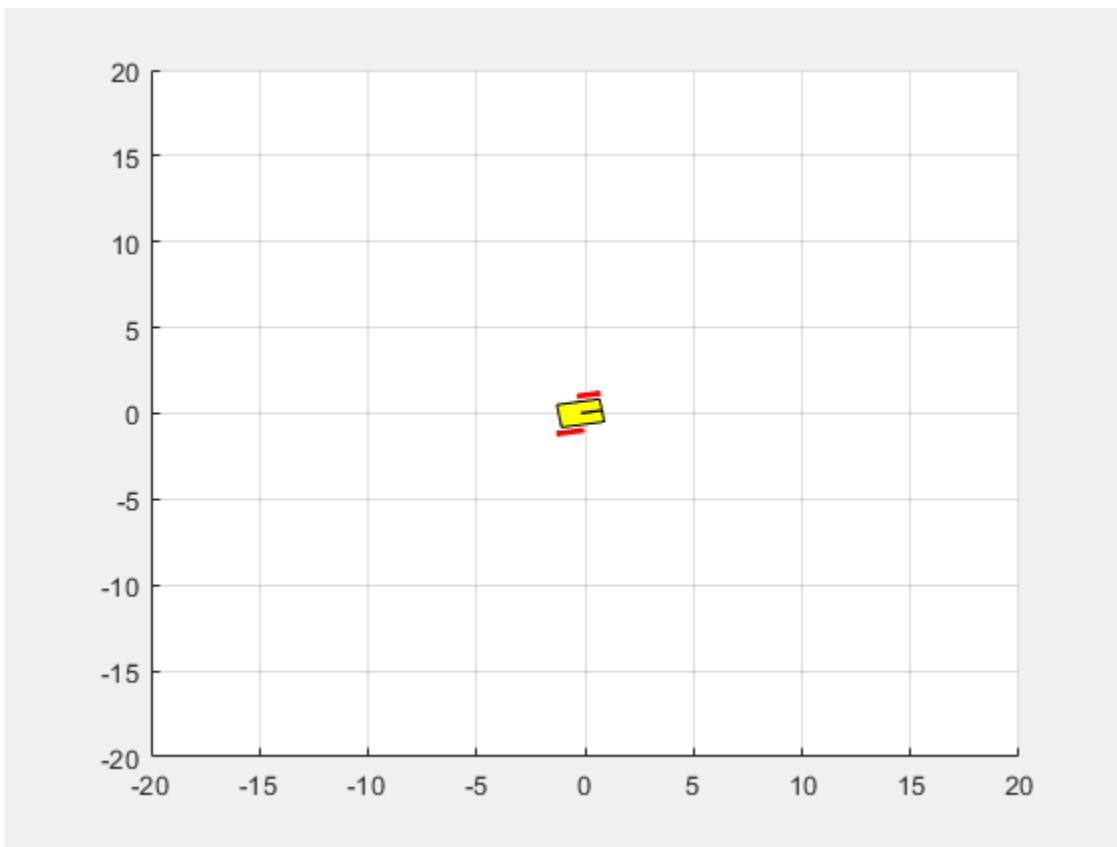
```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOptions);
else
    % Load pretrained agent for the example.
    load('FlyingRobotDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, simulate it within the pendulum environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',maxsteps);
experience = sim(env,agent,simOptions);
```



See Also

[rlDDPGAgent | train](#)

More About

- “Train Reinforcement Learning Agents” on page 5-2

Train PPO Agent to Land Rocket

This example shows how to train a proximal policy optimization (PPO) agent with a discrete action space to land a rocket on the ground. For more information on PPO agents, see “Proximal Policy Optimization Agents” on page 4-24.

Environment

The environment in this example is a 3-DOF rocket represented by a circular disc with mass. The rocket has two thrusters for forward and rotational motion. Gravity acts vertically downwards, and there are no aerodynamic drag forces. The training goal is to make the robot land on the ground at a specified location.

For this environment:

- Motion of the rocket is bounded in X (horizontal axis) from -100 to 100 meters and Y (vertical axis) from 0 to 120 meters.
- The goal position is at (0,0) meters and the goal orientation is 0 radians.
- The sample time is 0.1 seconds.
- The observations from the environment are the rocket position (x, y), orientation (θ), velocity (\dot{x}, \dot{y}), angular velocity ($\dot{\theta}$), and a sensor reading that detects rough landing (-1), soft landing (1) or airborne (0) condition. The observations are normalized between -1 and 1.
- At the beginning of every episode, the rocket starts from a random initial x position and orientation. The altitude is always reset to 100 meters.
- The reward r_t provided at the time step t is as follows.

$$\hat{d}_t = \sqrt{x_t^2 + y_t^2} / d_{\max}$$

$$\hat{v}_t = \sqrt{\dot{x}_t^2 + \dot{y}_t^2} / v_{\max}$$

$$r_1 = 1 - \left(\frac{\sqrt{\hat{d}_t} + \sqrt{\hat{v}_t}}{2} \right)$$

$$r_2 = \frac{1}{2} e^{-20\theta_t^2}$$

$$r_3 = 1 - \left(\frac{L_t + R_t}{20} \right)$$

$$r_4 = 10000[(y_t \leq 0) \& (\dot{y}_t \geq -0.5 \& \& |\dot{x}_t| \leq 0.5)]$$

$$r_t = r_1 + r_2 + r_3 + r_4$$

Here:

- x_t, y_t, \dot{x}_t , and \dot{y}_t are the positions and velocities of the rocket along the x and y axes.
- \hat{d}_t is the normalized distance of the rocket from the goal position.
- \hat{v}_t is the normalized speed of the rocket.
- d_{\max} and v_{\max} are the maximum distances and speeds within the state-space.
- θ_t is the orientation with respect to the vertical axis.

- L_t and R_t are the action values for the left and right thrusters.
- r_1 is a reward for minimizing distance and speed simultaneously.
- r_2 is a reward for minimizing the orientation of the rocket.
- r_3 is a reward for minimizing control effort.
- r_4 is a sparse reward for soft-landing with horizontal and vertical velocities less than 0.5 m/s.

Create MATLAB Environment

Create a MATLAB environment for the rocket lander using the `RocketLander` class.

```
env = RocketLander;
```

Obtain the observation and action information from the environment.

```
actionInfo = getActionInfo(env);
observationInfo = getObservationInfo(env);
numObs = observationInfo.Dimension(1);
numAct = numel(actionInfo.Elements);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create PPO Agent

The PPO agent in this example operates on a discrete action space. At every time step, the agent selects one of the following discrete action pairs.

L, L – do nothing
 L, M – fire right (med)
 L, H – fire right (high)
 M, L – fire left (med)
 M, M – fire left (med) + right (med)
 M, H – fire left (med) + right (high)
 H, L – fire left (high)
 H, M – fire left (high) + right (med)
 H, H – fire left (high) + right (high)

Here, $L = 0.0$, $M = 0.5$ and $H = 1.0$ are normalized thrust values for each thruster.

To estimate the policy and value function, the agent maintains function approximators for the actor and critic, which are modeled using deep neural networks. The training can be sensitive to the initial network weights and biases, and results can vary with different sets of values. In this example, the network weights are randomly initialized to small values. To load a set of predefined values instead, set `predefinedWeightsandBiases` to `true`.

```
criticLayerSizes = [200 100];
actorLayerSizes = [200 100];

predefinedWeightsandBiases = false;
if predefinedWeightsandBiases
    load('PredefinedWeightsAndBiases.mat');
```

```

else
    createNetworkWeights;
end

```

Create the critic deep neural network with six inputs and one output. The output of the critic network is the discounted long term reward for the input observations.

```

criticNetwork = [imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(criticLayerSizes(1), 'Name', 'CriticFC1', ...
        'Weights', weights.criticFC1, ...
        'Bias', bias.criticFC1)
    reluLayer('Name', 'CriticRelu1')
    fullyConnectedLayer(criticLayerSizes(2), 'Name', 'CriticFC2', ...
        'Weights', weights.criticFC2, ...
        'Bias', bias.criticFC2)
    reluLayer('Name', 'CriticRelu2')
    fullyConnectedLayer(1, 'Name', 'CriticOutput', ...
        'Weights', weights.criticOut, ...
        'Bias', bias.criticOut)];

```

Create the critic representation.

```

criticOpts = rlRepresentationOptions('LearnRate', 1e-3);
critic = rlValueRepresentation(criticNetwork, env.getObservationInfo, ...
    'Observation', {'observation'}, criticOpts);

```

Create the actor using a deep neural network with six inputs and two outputs. The outputs of the actor network are the probabilities of taking each possible action pair. Each action pair contains normalized action values for each thruster. The environment `step` function scales these values to determine the actual thrust values.

```

actorNetwork = [imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(actorLayerSizes(1), 'Name', 'ActorFC1', ...
        'Weights', weights.actorFC1, ...
        'Bias', bias.actorFC1)
    reluLayer('Name', 'ActorRelu1')
    fullyConnectedLayer(actorLayerSizes(2), 'Name', 'ActorFC2', ...
        'Weights', weights.actorFC2, ...
        'Bias', bias.actorFC2)
    reluLayer('Name', 'ActorRelu2')
    fullyConnectedLayer(numAct, 'Name', 'Action', ...
        'Weights', weights.actorOut, ...
        'Bias', bias.actorOut)
    softmaxLayer('Name', 'actionProbability')
];

```

Create the actor using a stochastic actor representation.

```

actorOptions = rlRepresentationOptions('LearnRate', 1e-3);
actor = rlStochasticActorRepresentation(actorNetwork, env.getObservationInfo, env.getActionInfo, ...
    'Observation', {'observation'}, actorOptions);

```

The PPO agent generates experiences for `ExperienceHorizon` number of steps or until the end of the episode. Then it trains on those experiences using mini-batches for a specified number of epochs.

- Set the `ExperienceHorizon` to 512 steps, which is slightly less than half the total simulation duration of 1200 steps.

- Use the "gae" method, which computes the advantage function based on the smoothed discounted sum of temporal differences.
- To improve training stability, use an objective function clip factor of 0.2.
- Discount factors that are close to 1 encourage long term rewards. Use a discount factor of 0.9995, which has a half-life of $\log(0.5)/\log(0.9995) \approx 1386$ steps.

Specify the agent options using `rlPPOAgentOptions`.

```
opt = rlPPOAgentOptions('ExperienceHorizon',512,...  
                         'ClipFactor',0.2,...  
                         'EntropyLossWeight',0.02,...  
                         'MiniBatchSize',64,...  
                         'NumEpoch',3,...  
                         'AdvantageEstimateMethod','gae',...  
                         'GAEFactor',0.95,...  
                         'SampleTime',env.Ts,...  
                         'DiscountFactor',0.9995);
```

Create the PPO agent.

```
agent = rlPPOAgent(actor,critic,opt);
```

Train Agent

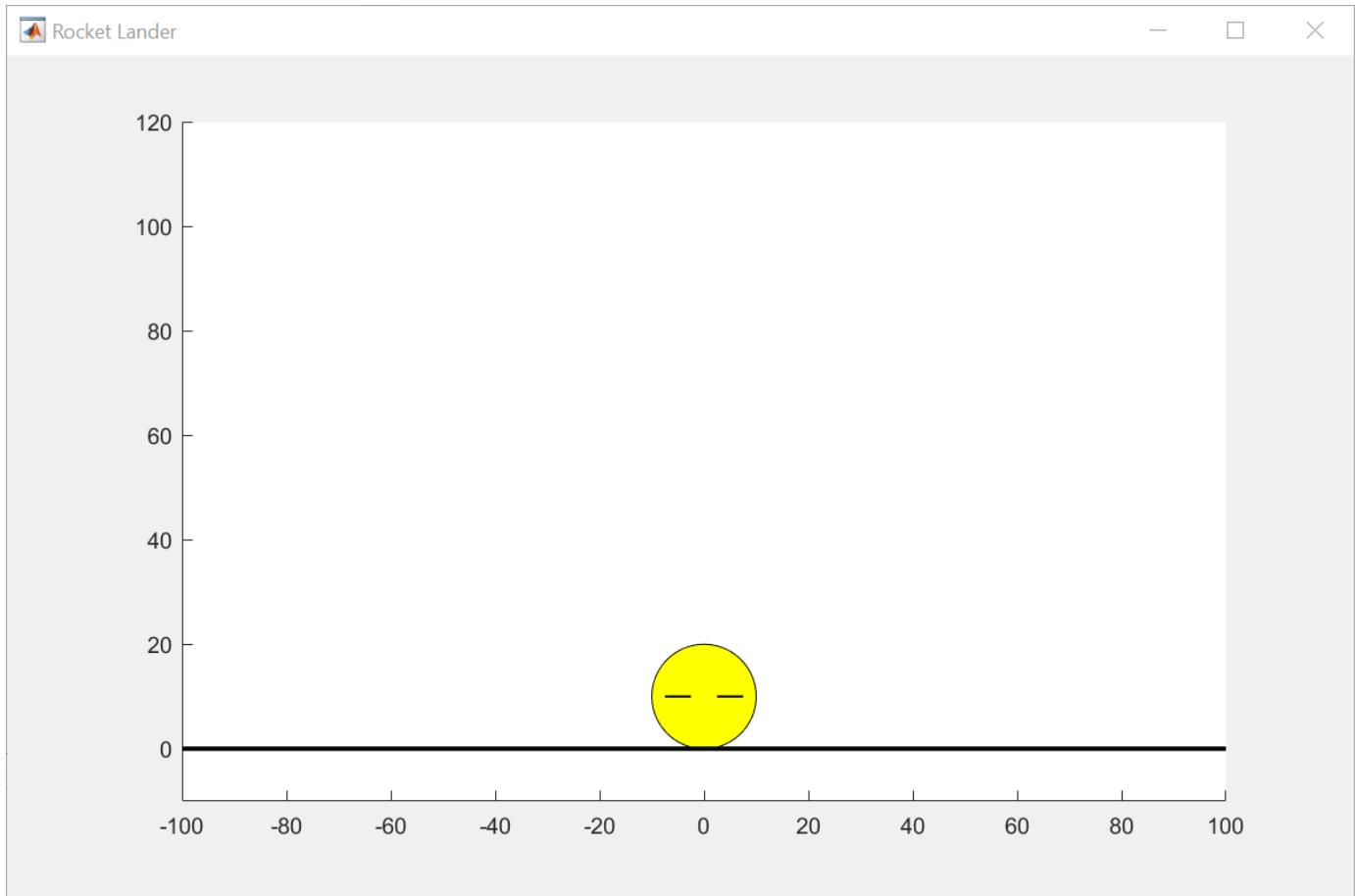
To train the PPO agent, specify the following training options.

- Run the training for at most 20000 episodes, with each episode lasting at most 1200 time steps.
- Stop the training when the average reward over 100 consecutive episodes is 10000. This value indicates that the agent is consistently receiving the reward for soft landing.
- Save a copy of the agent for each episode where the episode reward is 11000 or more.

```
trainOpts = rlTrainingOptions(...  
                           'MaxEpisodes',20000,...  
                           'MaxStepsPerEpisode',1200,...  
                           'Verbose',false,...  
                           'Plots','training-progress',...  
                           'StopTrainingCriteria','AverageReward',...  
                           'StopTrainingValue',10000,...  
                           'ScoreAveragingWindowLength',100,...  
                           'SaveAgentCriteria',"EpisodeReward",...  
                           'SaveAgentValue',11000);
```

Plot the rocket lander environment to visualize the training or simulation.

```
plot(env)
```



Train the agent using the `train` function. Due to the complexity of the environment, this process is computationally intensive and takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`.

```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained parameters for the example.
    load('RocketLanderPPOAgentParams.mat');
    loadPretrainedParams(agent,actorParams,criticParams)
end

ans =
    rlPPOAgent with properties:
        AgentOptions: [1x1 rl.option.rlPPOAgentOptions]
```

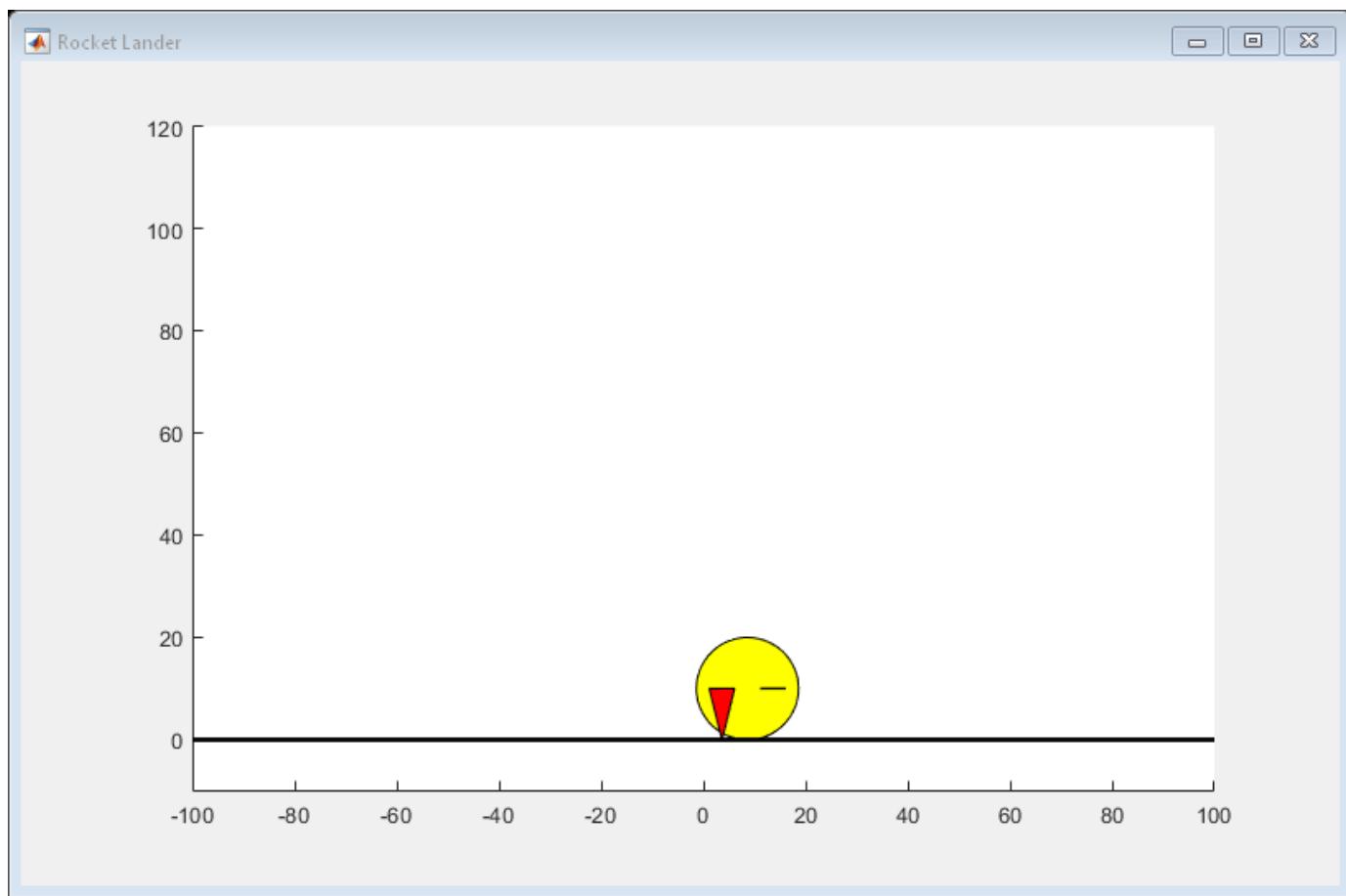
5 Train and Validate Agents



Simulate

Simulate the trained agent within the environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
Tf = 120; % Total simulation length (seconds)
simOptions = rlSimulationOptions('MaxSteps',ceil(Tf/env.Ts));
experience = sim(env,agent,simOptions);
```



Local Function

Function to update agent with pretrained parameters

```
function agent = loadPretrainedParams(agent,actorParams,criticParams)
% Set actor parameters.
actor = getActor(agent);
pretrainedActor = setLearnableParameters(actor,actorParams);
% Set critic parameters.
critic = getCritic(agent);
pretrainedCritic = setLearnableParameters(critic,criticParams);
% Set actor and critic in agent.
agent = setActor(agent,pretrainedActor);
agent = setCritic(agent,pretrainedCritic);
end
```

See Also

[rlPPOAgent](#) | [train](#)

More About

- “Train Reinforcement Learning Agents” on page 5-2

Train DDPG Agent for Adaptive Cruise Control

This example shows how to train a deep deterministic policy gradient (DDPG) agent for adaptive cruise control (ACC) in Simulink®. For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14.

Simulink Model

The reinforcement learning environment for this example is the simple longitudinal dynamics for an ego car and lead car. The training goal is to make the ego car travel at a set velocity while maintaining a safe distance from lead car by controlling longitudinal acceleration and braking. This example uses the same vehicle model as the “Adaptive Cruise Control System Using Model Predictive Control” (Model Predictive Control Toolbox) example.

Specify the initial position and velocity for the two vehicles.

```
x0_lead = 50; % initial position for lead car (m)
v0_lead = 25; % initial velocity for lead car (m/s)
x0_ego = 10; % initial position for ego car (m)
v0_ego = 20; % initial velocity for ego car (m/s)
```

Specify standstill default spacing (m), time gap (s) and driver-set velocity (m/s).

```
D_default = 10;
t_gap = 1.4;
v_set = 30;
```

Considering the physical limitations of the vehicle dynamics, the acceleration is constrained to the range [-3, 2] (m/s²).

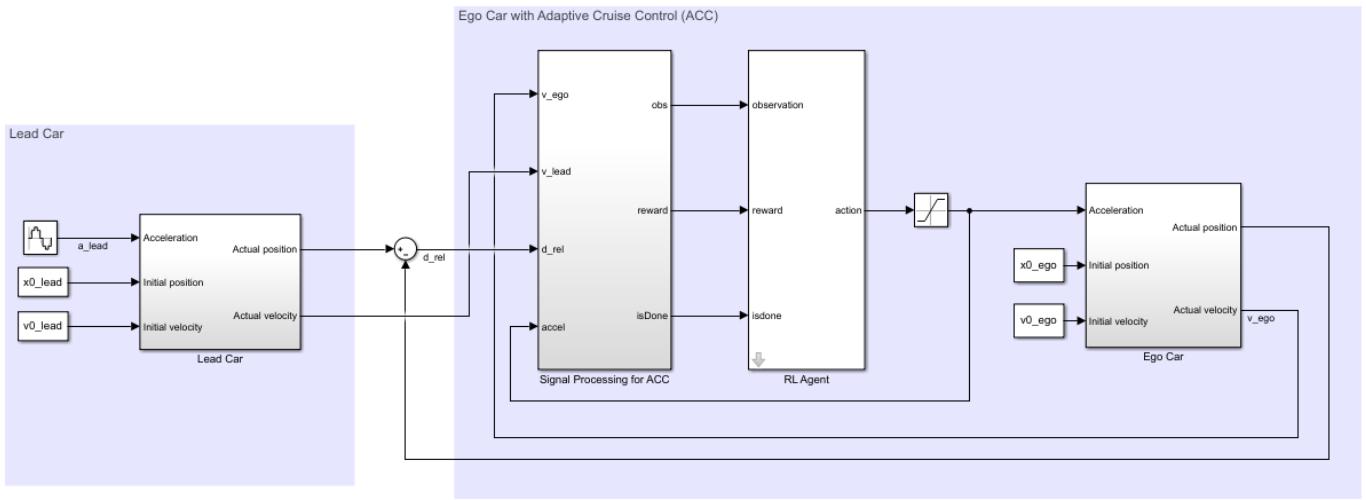
```
amin_ego = -3;
amax_ego = 2;
```

Define the sample time Ts and simulation duration Tf in seconds.

```
Ts = 0.1;
Tf = 60;
```

Open the model.

```
mdl = 'rlACCMdl';
open_system(mdl)
agentblk = [mdl '/RL Agent'];
```



Copyright 2018 The MathWorks, Inc.

For this model:

- The acceleration action signal from the agent to the environment is from -3 to 2 m/s².
- The reference velocity for ego car V_{ref} is defined as follows. If relative distance is less than safe distance, ego car tracks the minimum of lead car velocity and driver-set velocity. In this manner, the ego car maintains some distance from lead car. If the relative distance is greater than safe distance, the ego car tracks driver-set velocity. In this example, safe distance is defined as a linear function of ego car longitudinal velocity V ; that is, $t_{gap} * V + D_{default}$. The safe distance determines the reference tracking velocity for the ego car.
- The observations from the environment are the velocity error $e = V_{ref} - V_{ego}$, its integral $\int e$, and the ego car longitudinal velocity V .
- The simulation is terminated when longitudinal velocity of the ego car is less than 0, or the relative distance between the lead car and ego car becomes less than 0.
- The reward r_t , provided at every time step t , is:

$$r_t = -(0.1e_t^2 + u_{t-1}^2) + M_t$$

where u_{t-1} is the control input from the previous time step. The logical value $M_t = 1$ if velocity error $e_t^2 \leq 0.25$, otherwise $M_t = 0$.

Create Environment Interface

Create a reinforcement learning environment interface for the model.

```
% create the observation info
observationInfo = rlNumericSpec([3 1], 'LowerLimit', -inf*ones(3,1), 'UpperLimit', inf*ones(3,1));
observationInfo.Name = 'observations';
observationInfo.Description = 'information on velocity error and ego velocity';
% action Info
actionInfo = rlNumericSpec([1 1], 'LowerLimit', -3, 'UpperLimit', 2);
actionInfo.Name = 'acceleration';
% define environment
env = rlSimulinkEnv(mdl, agentblk, observationInfo, actionInfo);
```

To define the initial condition for the position of the lead car, specify an environment reset function using an anonymous function handle.

```
% randomize initial positions of lead car  
env.ResetFcn = @(in)localResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng('default')
```

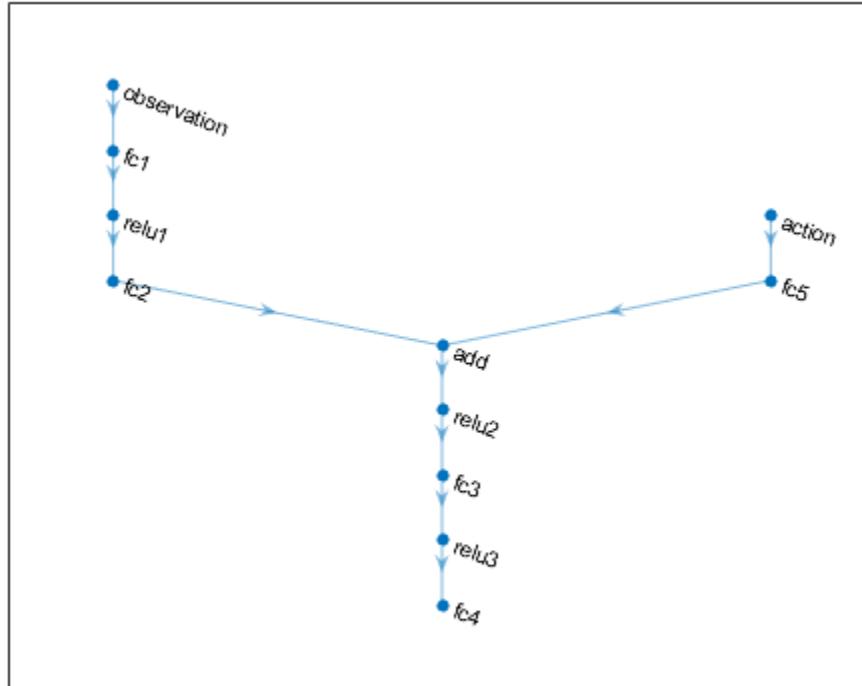
Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
L = 48; % number of neurons  
statePath = [  
    imageInputLayer([3 1 1], 'Normalization', 'none', 'Name', 'observation')  
    fullyConnectedLayer(L, 'Name', 'fc1')  
    reluLayer('Name', 'relu1')  
    fullyConnectedLayer(L, 'Name', 'fc2')  
    additionLayer(2, 'Name', 'add')  
    reluLayer('Name', 'relu2')  
    fullyConnectedLayer(L, 'Name', 'fc3')  
    reluLayer('Name', 'relu3')  
    fullyConnectedLayer(1, 'Name', 'fc4')];  
  
actionPath = [  
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')  
    fullyConnectedLayer(L, 'Name', 'fc5')];  
  
criticNetwork = layerGraph(statePath);  
criticNetwork = addLayers(criticNetwork, actionPath);  
  
criticNetwork = connectLayers(criticNetwork, 'fc5', 'add/in2');
```

View the critic network configuration.

```
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate',1e-3,'GradientThreshold',1,'L2Regularization',1);
```

Create the critic representation using the specified neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,observationInfo,actionInfo,...  
    'Observation',{'observation'},'Action',{'action'},criticOptions);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input, the observation, and one output, the action.

Construct the actor similarly to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [  
    imageInputLayer([3 1 1],'Normalization','none','Name','observation')  
    fullyConnectedLayer(L,'Name','fc1')  
    reluLayer('Name','relu1')  
    fullyConnectedLayer(L,'Name','fc2')  
    reluLayer('Name','relu2')  
    fullyConnectedLayer(L,'Name','fc3')  
    reluLayer('Name','relu3')  
    fullyConnectedLayer(1,'Name','fc4')]
```

```
tanhLayer('Name','tanh1')
scalingLayer('Name','ActorScaling1','Scale',2.5,'Bias',-0.5)];

actorOptions = rlRepresentationOptions('LearnRate',1e-4,'GradientThreshold',1,'L2RegularizationFactor',1);
actor = rlDeterministicActorRepresentation(actorNetwork,observationInfo,actionInfo,...)
    'Observation',{['observation']}, 'Action',{['ActorScaling1']}, actorOptions);
```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```
agentOptions = rlDDPGAgentOptions(... ...
    'SampleTime',Ts, ...
    'TargetSmoothFactor',1e-3, ...
    'ExperienceBufferLength',1e6, ...
    'DiscountFactor',0.99, ...
    'MiniBatchSize',64);
agentOptions.NoiseOptions.Variance = 0.6;
agentOptions.NoiseOptions.VarianceDecayRate = 1e-5;
```

Then, create the DDPG agent using the specified actor representation, critic representation, and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training episode for at most 5000 episodes, with each episode lasting at most 600 time steps.
- Display the training progress in the Episode Manager dialog box.
- Stop training when the agent receives an episode reward greater than 260.

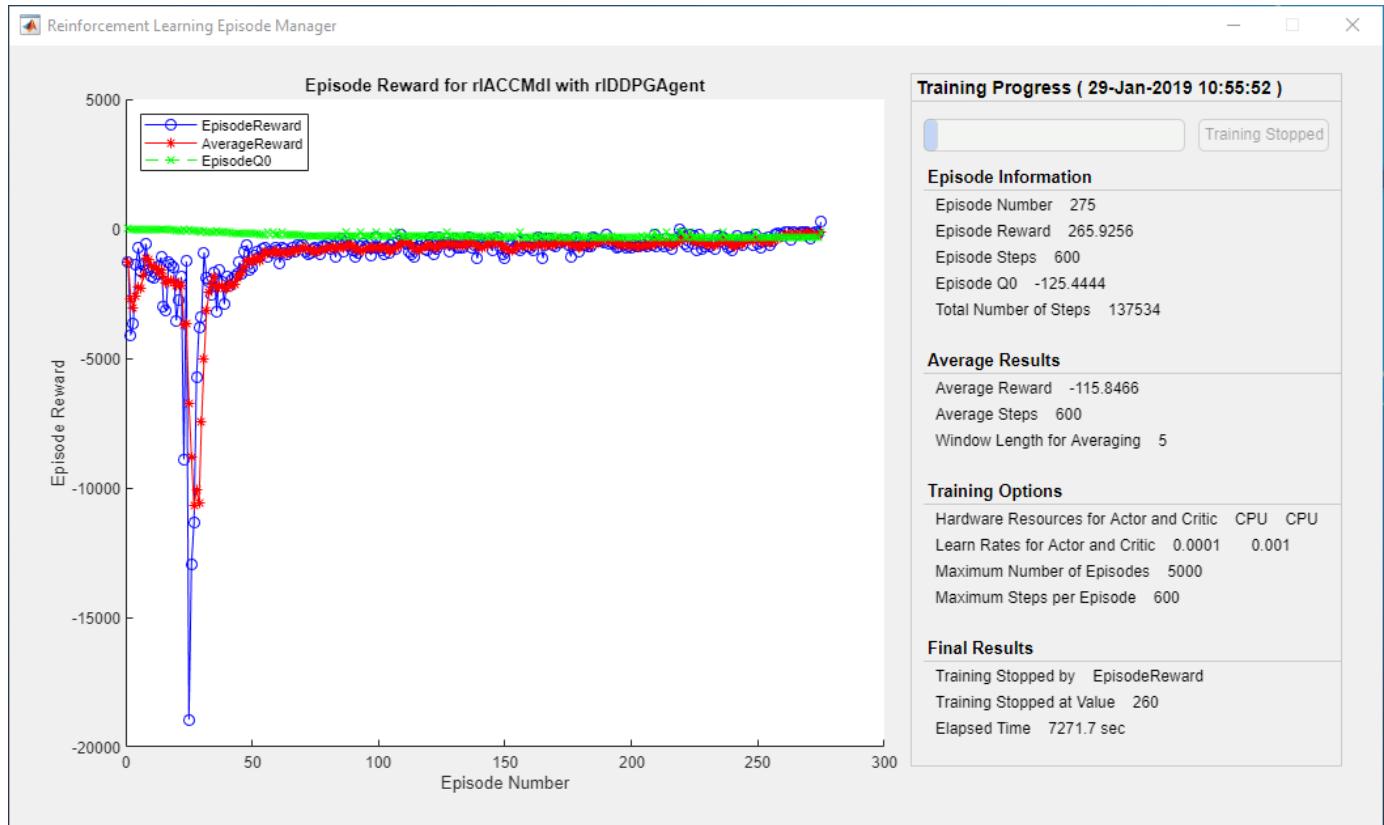
For more information, see `rlTrainingOptions`.

```
maxepisodes = 5000;
maxsteps = ceil(Tf/Ts);
trainingOpts = rlTrainingOptions(... ...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','EpisodeReward', ...
    'StopTrainingValue',260);
```

Train the agent using the `train` function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOpts);
else
    % Load pretrained agent for the example.
    load('SimulinkACCDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, uncomment the following commands and simulate it within the Simulink environment. For more information on agent simulation, see [rlSimulationOptions](#) and [sim](#).

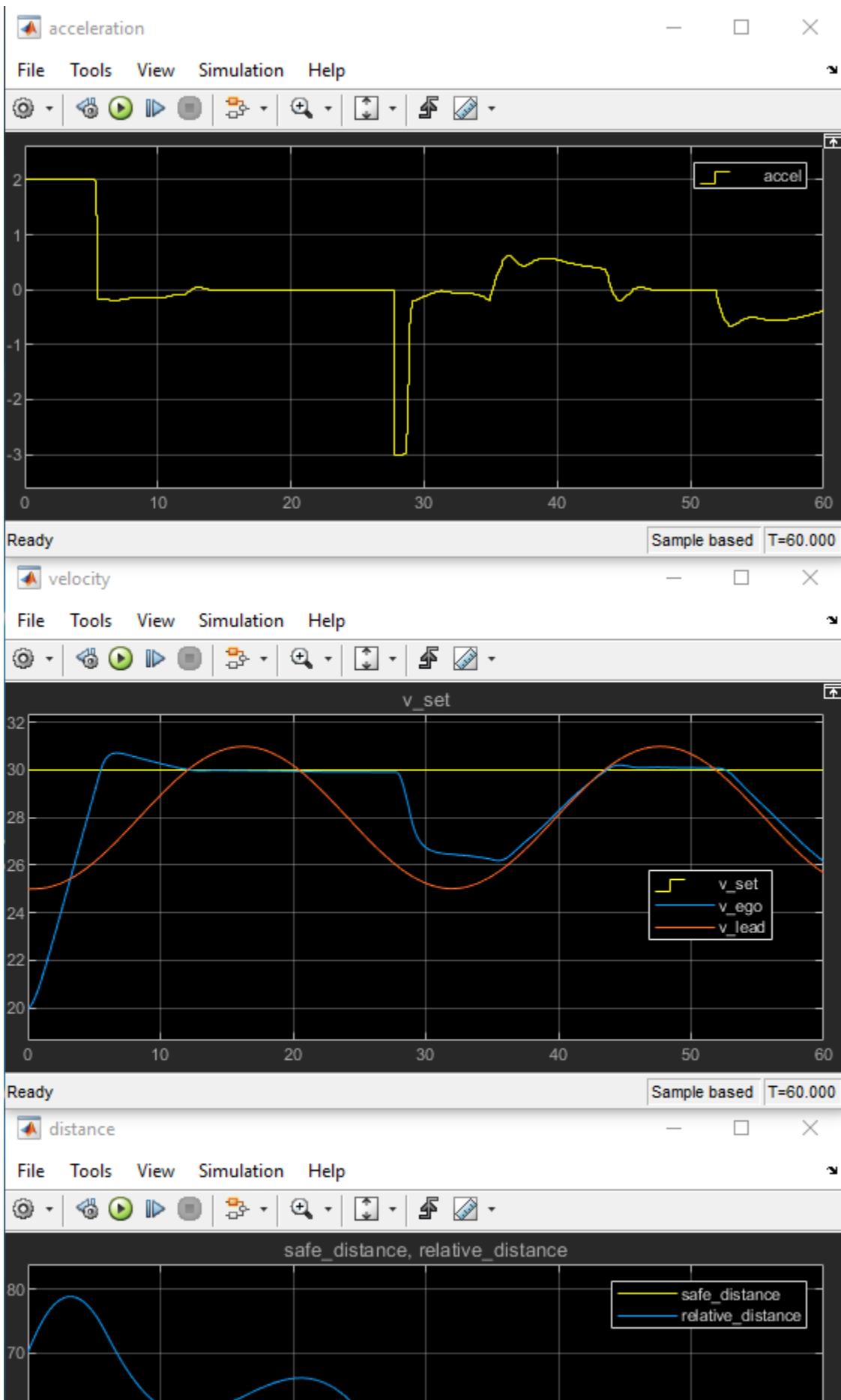
```
% simOptions = rlSimulationOptions('MaxSteps',maxsteps);
% experience = sim(env,agent,simOptions);
```

To demonstrate the trained agent using deterministic initial conditions, simulate the model in Simulink.

```
x0_lead = 80;
sim(mdl)
```

The following plots show the simulation results when lead car is 70 (m) ahead of ego car.

- In the first 28 seconds, relative distance is greater than safe distance (bottom plot), therefore the ego car tracks set velocity (middle plot). To speed up and reach the set velocity, acceleration is positive (top plot).
- From 28 to 60 seconds, relative distance is less than safe distance (bottom plot), therefore the ego car tracks the minimum of the lead velocity and set velocity. From 28 to 36 seconds, the lead velocity is less than the set velocity (middle plot). To slow down and track the lead car velocity, acceleration is negative (top plot). From 36 to 60 seconds, ego car adjusts its acceleration to track the reference velocity closely (middle plot). Within this time interval, the ego car tracks the set velocity from 43 to 52 seconds and tracks lead velocity from 36 to 43 seconds and 52 to 60 seconds.



Close the Simulink model.

```
bdclose(mdl)
```

Local Function

```
function in = localResetFcn(in)
% reset initial position of lead car
in = setVariable(in,'x0_lead',40+randi(60,1,1));
end
```

See Also

[train](#)

More About

- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Train DQN Agent for Lane Keeping Assist

This example shows how to train a Deep Q-Learning Network (DQN) agent for lane keeping assist (LKA) in Simulink®. For more information on DQN agents, see “Deep Q-Network Agents” on page 4-8.

Simulink Model for Ego Car

The reinforcement learning environment for this example is a simple bicycle model for ego vehicle dynamics. The training goal is to keep the ego vehicle traveling along the centerline of the lanes by adjusting the front steering angle. This example uses the same vehicle model as in “Lane Keeping Assist System Using Model Predictive Control” (Model Predictive Control Toolbox) example.

```
m = 1575; % total vehicle mass (kg)
Iz = 2875; % yaw moment of inertia (mNs^2)
lf = 1.2; % longitudinal distance from center of gravity to front tires (m)
lr = 1.6; % longitudinal distance from center of gravity to rear tires (m)
Cf = 19000; % cornering stiffness of front tires (N/rad)
Cr = 33000; % cornering stiffness of rear tires (N/rad)
Vx = 15; % longitudinal velocity (m/s)
```

Define the sample time, T_s , and simulation duration, T , in seconds.

```
Ts = 0.1;
T = 15;
```

The output of the LKA system is the front steering angle of the ego car. Considering the physical limitations of the ego car, the steering angle is constrained to the range [-0.5,0.5] rad.

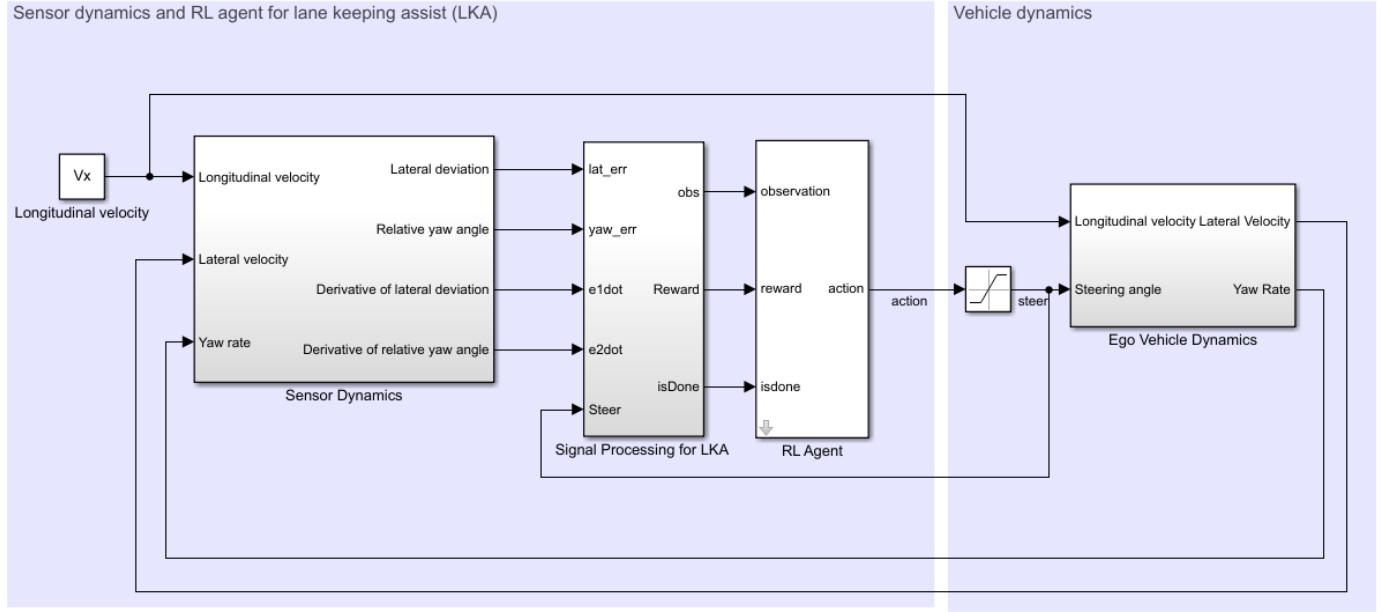
```
u_min = -0.5;
u_max = 0.5;
```

The curvature of the road is defined by a constant $0.001(m^{-1})$. The initial value for lateral deviation is 0.2 m and the initial value for relative yaw angle is -0.1 rad.

```
rho = 0.001;
e1_initial = 0.2;
e2_initial = -0.1;
```

Open the model.

```
mdl = 'rLLKAMdl';
open_system(mdl);
agentblk = [mdl '/RL Agent'];
```



Copyright 2018 The MathWorks, Inc.

For this model:

- The steering-angle action signal from the agent to the environment is from -15 deg to 15 deg.
- The observations from the environment are the lateral deviation e_1 , relative yaw angle e_2 , their derivatives \dot{e}_1 and \dot{e}_2 , and their integrals $\int e_1$ and $\int e_2$.
- The simulation is terminated when lateral deviation $|e_1| > 1$.
- The reward r_t , provided at every time step t , is:

$$r_t = -(10e_1^2 + 5e_2^2 + 2u^2 + 5\dot{e}_1^2 + 5\dot{e}_2^2)$$

where u is the control input from the previous time step $t - 1$.

Create Environment Interface

Create a reinforcement learning environment interface for the ego vehicle. To do so, first create the observation and action specifications.

```
% create observation info
observationInfo = rlNumericSpec([6 1], 'LowerLimit', -inf*ones(6,1), 'UpperLimit', inf*ones(6,1));
observationInfo.Name = 'observations';
observationInfo.Description = 'information on lateral deviation and relative yaw angle';
% create action Info
actionInfo = rlFiniteSetSpec((-15:15)*pi/180);
actionInfo.Name = 'steering';
% define environment
env = rlSimulinkEnv(mdl, agentblk, observationInfo, actionInfo);
```

The interface has a discrete action space where the agent can apply one of 31 possible steering angles from -15 degrees to 15 degrees.

To define the initial condition for lateral deviation and relative yaw angle, specify an environment reset function using an anonymous function handle.

```
% randomize initial values for lateral deviation and relative yaw angle
env.ResetFcn = @(in)localResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DQN agent

A DQN agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

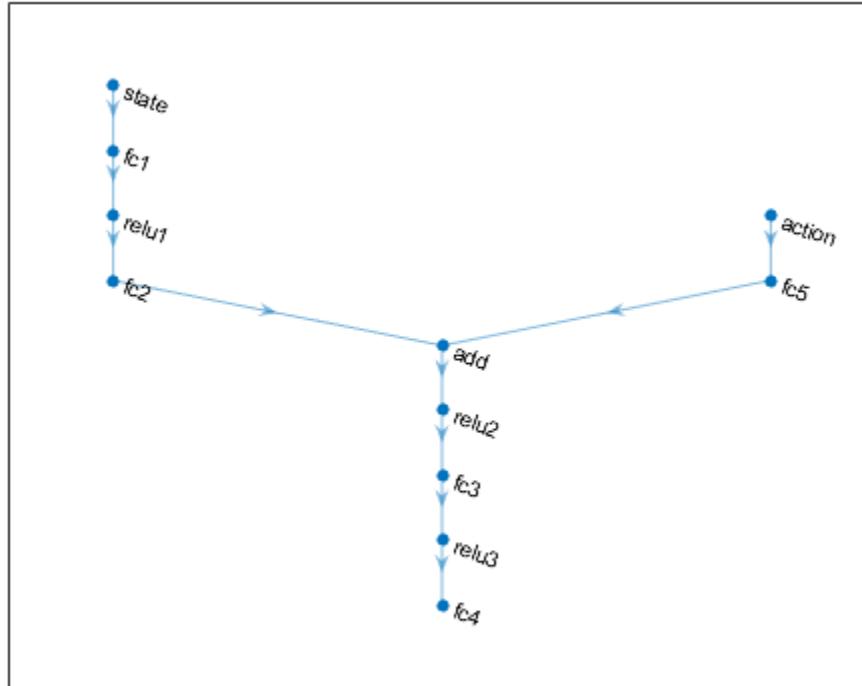
```
L = 24; % number of neurons
statePath = [
    imageInputLayer([6 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(L, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(L, 'Name', 'fc2')
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(L, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(1, 'Name', 'fc4')];

actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(L, 'Name', 'fc5')];

criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = connectLayers(criticNetwork, 'fc5', 'add/in2');
```

View the critic network configuration.

```
figure
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate',1e-3,'GradientThreshold',1,'L2Regularization',1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation information for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,observationInfo,actionInfo,...  
    'Observation',{'state'},'Action',{'action'},criticOptions);
```

To create the DQN agent, first specify the DQN agent options using `rldQNAgentOptions`.

```
agentOptions = rldQNAgentOptions(...  
    'SampleTime',Ts,...  
    'UseDoubleDQN',true,...  
    'TargetSmoothFactor',1e-3,...  
    'DiscountFactor',0.99,...  
    'ExperienceBufferLength',1e6,...  
    'MiniBatchSize',64);
```

Then, create the DQN agent using the specified critic representation and agent options. For more information, see `rldQNAgent`.

```
agent = rldQNAgent(critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training episode for at most 5000 episodes, with each episode lasting at most $\text{ceil}(T/T_s)$ time steps.
- Display the training progress in the Episode Manager dialog box.
- Stop training when the episode reward reaches -1.
- Save a copy of the agent for each episode where the cumulative reward is greater than -2.5.

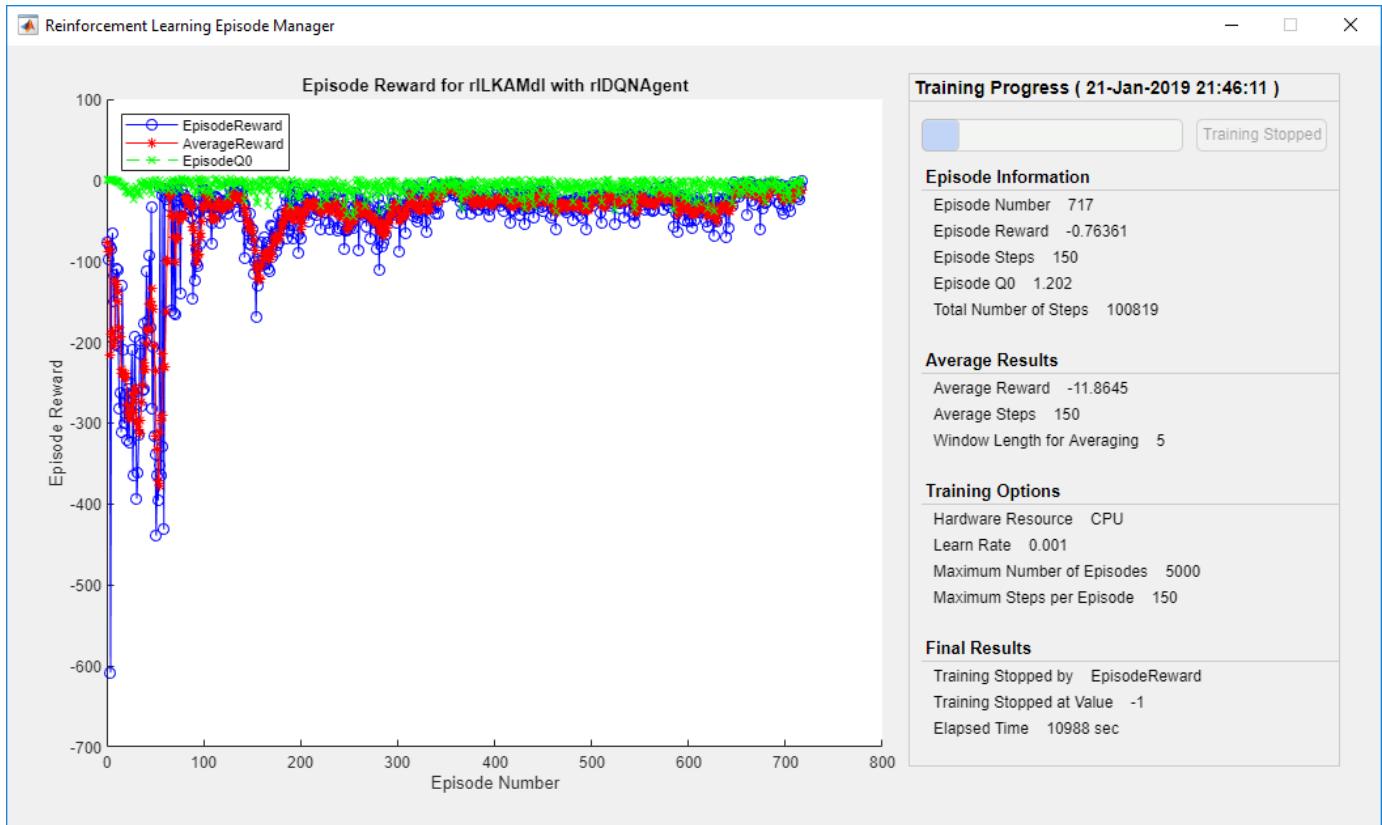
For more information, see [rlTrainingOptions](#).

```
maxepisodes = 5000;
maxsteps = ceil(T/Ts);
trainingOpts = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','EpisodeReward', ...
    'StopTrainingValue',-1, ...
    'SaveAgentCriteria','EpisodeReward', ...
    'SaveAgentValue',-2.5);
```

Train the agent using the `train` function. This is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOpts);
else
    % Load pretrained agent for the example.
    load('SimulinkLKADQN.mat','agent')
end
```



Simulate DQN Agent

To validate the performance of the trained agent, uncomment the following two lines and simulate it within the environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

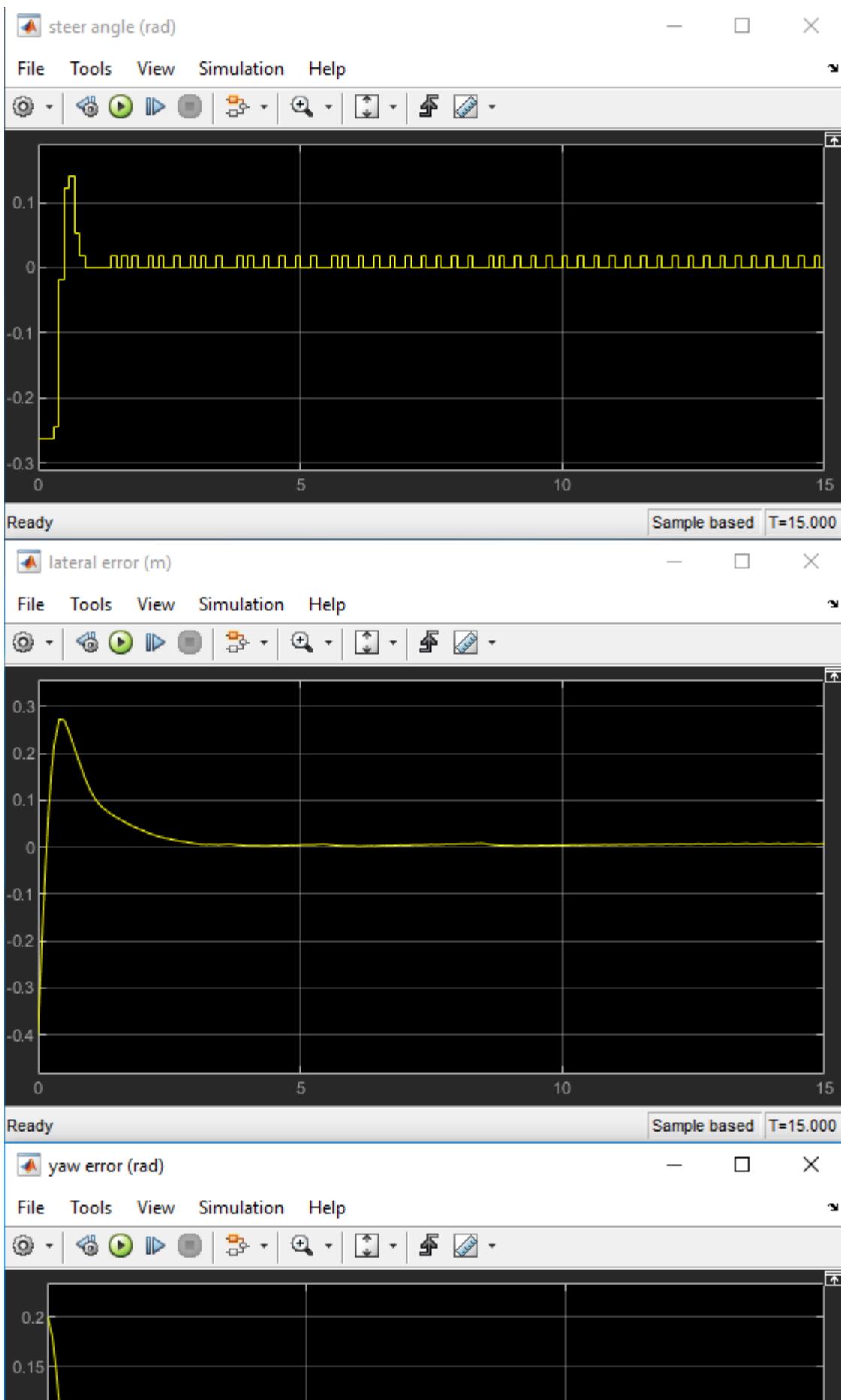
```
% simOptions = rlSimulationOptions('MaxSteps',maxsteps);
% experience = sim(env,agent,simOptions);
```

To demonstrate the trained agent on deterministic initial conditions, simulate the model in Simulink.

```
e1_initial = -0.4;
e2_initial = 0.2;
sim(mdl)
```

As shown below, the lateral error (middle plot) and relative yaw angle (bottom plot) are both driven to zero. The vehicle starts from off centerline (-0.4 m) and non-zero yaw angle error (0.2 rad). The lane keeping assist makes the ego car traveling along the centerline around 2.5 seconds. The steering angle (top plot) shows that the controller reaches steady-state after 2 seconds.

5 Train and Validate Agents



Close the Simulink model.

```
bdclose(mdl)
```

Local Function

```
function in = localResetFcn(in)
% reset
in = setVariable(in,'e1_initial', 0.5*(-1+2*rand)); % random value for lateral deviation
in = setVariable(in,'e2_initial', 0.1*(-1+2*rand)); % random value for relative yaw angle
end
```

See Also

`train`

More About

- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Train DDPG Agent for Path Following Control

This example shows how to train a deep deterministic policy gradient (DDPG) agent for path-following control (PFC) in Simulink®. For more information on DDPG agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14.

Simulink Model

The reinforcement learning environment for this example is a simple bicycle model for ego car and a simple longitudinal model for lead car. The training goal is to make the ego car travel at a set velocity while maintaining a safe distance from lead car by controlling longitudinal acceleration and braking, while also keeping the ego car travelling along the centerline of its lane by controlling the front steering angle. For more information on PFC, see Path Following Control System. The ego car dynamics are specified by the following parameters.

```
m = 1600; % total vehicle mass (kg)
Iz = 2875; % yaw moment of inertia (mNs^2)
lf = 1.4; % longitudinal distance from center of gravity to front tires (m)
lr = 1.6; % longitudinal distance from center of gravity to rear tires (m)
Cf = 19000; % cornering stiffness of front tires (N/rad)
Cr = 33000; % cornering stiffness of rear tires (N/rad)
tau = 0.5; % longitudinal time constant
```

Specify the initial position and velocity for the two vehicles.

```
x0_lead = 50; % initial position for lead car (m)
v0_lead = 24; % initial velocity for lead car (m/s)
x0_ego = 10; % initial position for ego car (m)
v0_ego = 18; % initial velocity for ego car (m/s)
```

Specify standstill default spacing (m), time gap (s) and driver-set velocity (m/s).

```
D_default = 10;
t_gap = 1.4;
v_set = 28;
```

Considering the physical limitations of the vehicle dynamics, the acceleration is constrained to the range [-3, 2] (m/s²), and steering angle is constrained to be [-0.5, 0.5] (rad).

```
a_min_ego = -3;
a_max_ego = 2;
u_min_ego = -0.5;
u_max_ego = 0.5;
```

The curvature of the road is defined by a constant $0.001(m^{-1})$. The initial value for lateral deviation is 0.2 m and the initial value for relative yaw angle is -0.1 rad.

```
rho = 0.001;
e1_initial = 0.2;
e2_initial = -0.1;
```

Define the sample time, Ts, and simulation duration, Tf, in seconds.

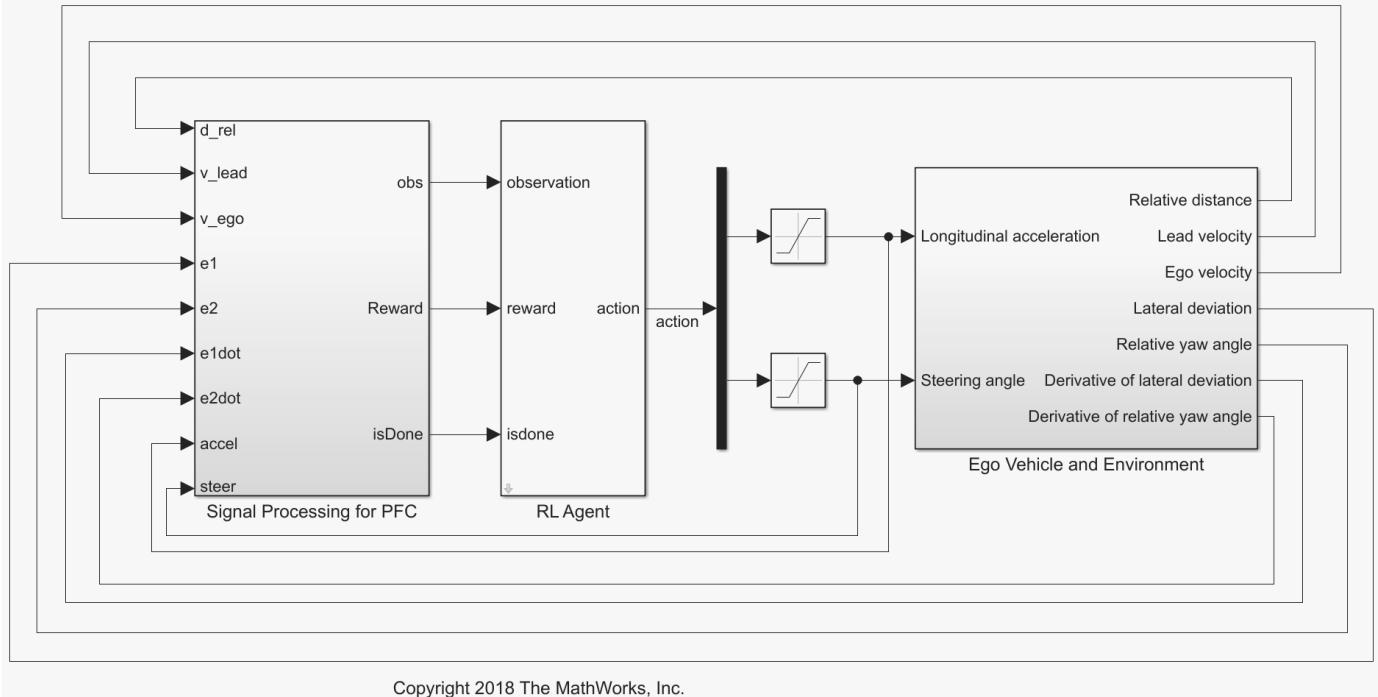
```
Ts = 0.1;
Tf = 60;
```

Open the model.

```

mdl = 'rlPFCMdl';
open_system(mdl)
agentblk = [mdl '/RL Agent'];

```



Copyright 2018 The MathWorks, Inc.

For this model:

- The action signal consists of acceleration and steering angle actions. The acceleration action signal takes value between -3 and 2 (m/s^2). The steering action signal takes value between -15 degrees (-0.2618 rad) to 15 degrees (0.2618 rad).
- The reference velocity for ego car V_{ref} is defined as follows. If relative distance is less than safe distance, ego car tracks the minimum of lead car velocity and driver-set velocity. In this manner, ego car maintains some distance from lead car. If relative distance is greater than safe distance, ego car tracks driver-set velocity. In this example, safe distance is defined as a linear function of ego car longitudinal velocity V , that is, $t_{gap} * V + D_{default}$. The safe distance determines the tracking velocity for the ego car.
- The observations from the environment contain the longitudinal measurements: the velocity error $e_V = V_{ref} - V_{ego}$, its integral $\int e$ and ego car longitudinal velocity V . In addition, the observations contain the lateral measurements: the lateral deviation e_1 , relative yaw angle e_2 , their derivatives \dot{e}_1 and \dot{e}_2 , and their integrals $\int e_1$ and $\int e_2$.
- The simulation is terminated when lateral deviation $|e_1| > 1$ or longitudinal velocity $V_{ego} < 0.5$ or relative distance between lead car and ego car $D_{rel} < 0$.
- The reward r_t , provided at every time step t , is:

$$r_t = -(100e_1^2 + 500u_{t-1}^2 + 10e_V^2 + 100a_{t-1}^2) \times 1e^{-3} - 10F_t + 2H_t + M_t$$

where u_{t-1} is the steering input from the previous time step $t-1$, a_{t-1} is the acceleration input from the previous time step. The three logical values are: $F_t = 1$ if simulation is terminated, otherwise

$F_t = 0$; $H_t = 1$ if lateral error $e_l^2 < 0.01$, otherwise $H_t = 0$; $M_t = 1$ if velocity error $e_V^2 < 1$, otherwise $M_t = 0$. The three logical terms in the reward encourage the agent to make both lateral error and velocity error small, in the meantime, penalize the agent if the simulation is terminated early.

Create Environment Interface

Create an environment interface for the Simulink model.

```
% create the observation info
observationInfo = rlNumericSpec([9 1], 'LowerLimit', -inf*ones(9,1), 'UpperLimit', inf*ones(9,1));
observationInfo.Name = 'observations';
% action Info
actionInfo = rlNumericSpec([2 1], 'LowerLimit', [-3;-0.2618], 'UpperLimit', [2;0.2618]);
actionInfo.Name = 'accel;steer';
% define environment
env = rlSimulinkEnv(mdl,agentblk,observationInfo,actionInfo);
```

To define the initial conditions, specify an environment reset function using an anonymous function handle.

```
% randomize initial positions of lead car, lateral deviation and relative
% yaw angle
env.ResetFcn = @(in)localResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DDPG agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
L = 100; % number of neurons
statePath = [
    imageInputLayer([9 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(L, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(L, 'Name', 'fc2')
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(L, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(1, 'Name', 'fc4')];

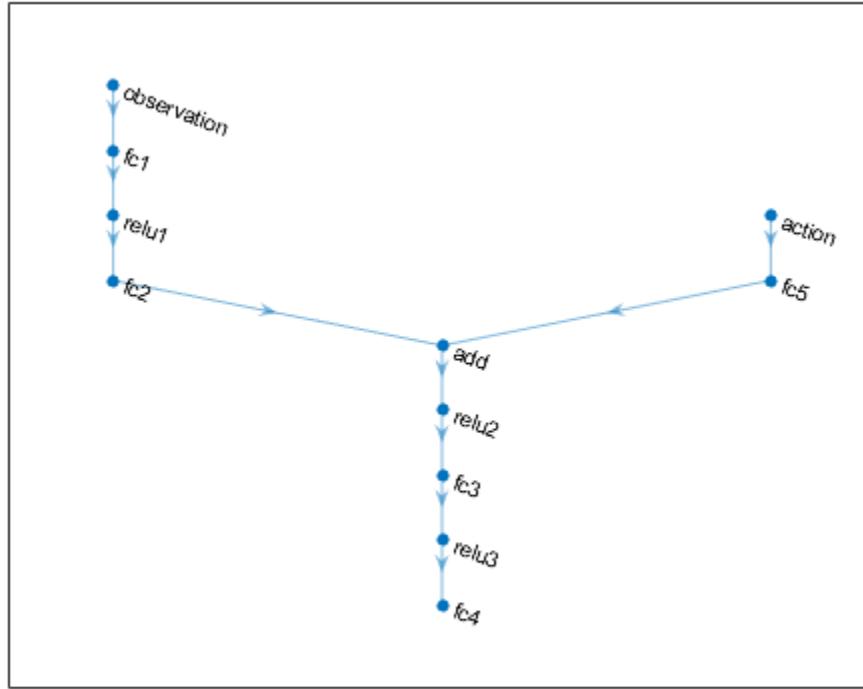
actionPath = [
    imageInputLayer([2 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(L, 'Name', 'fc5')];

criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork,actionPath);

criticNetwork = connectLayers(criticNetwork, 'fc5', 'add/in2');
```

View the critic network configuration.

```
figure
plot(criticNetwork)
```



Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOptions = rlRepresentationOptions('LearnRate',1e-3,'GradientThreshold',1,'L2Regularization',1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,observationInfo,actionInfo,...  
    'Observation',{'observation'},'Action',{'action'},criticOptions);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input, the observation, and one output, the action.

Construct the actor similarly to the critic. For more information, see `rlDeterministicActorRepresentation`.

```
actorNetwork = [  
    imageInputLayer([9 1 1],'Normalization','none','Name','observation')  
    fullyConnectedLayer(L,'Name','fc1')  
    reluLayer('Name','relu1')  
    fullyConnectedLayer(L,'Name','fc2')  
    reluLayer('Name','relu2')]
```

```

fullyConnectedLayer(L,'Name','fc3')
reluLayer('Name','relu3')
fullyConnectedLayer(2,'Name','fc4')
tanhLayer('Name','tanh1')
scalingLayer('Name','ActorScaling1','Scale',reshape([2.5;0.2618],[1,1,2]),'Bias',reshape([-0
actorOptions = rlRepresentationOptions('LearnRate',1e-4,'GradientThreshold',1,'L2RegularizationFa
actor = rlDeterministicActorRepresentation(actorNetwork,observationInfo,actionInfo, ...
'Observation',{ 'observation'},'Action',{ 'ActorScaling1'}},actorOptions);

```

To create the DDPG agent, first specify the DDPG agent options using `rlDDPGAgentOptions`.

```

agentOptions = rlDDPGAgentOptions(...
    'SampleTime',Ts, ...
    'TargetSmoothFactor',1e-3, ...
    'ExperienceBufferLength',1e6, ...
    'DiscountFactor',0.99, ...
    'MiniBatchSize',64);
agentOptions.NoiseOptions.Variance = [0.6;0.1];
agentOptions.NoiseOptions.VarianceDecayRate = 1e-5;

```

Then, create the DDPG agent using the specified actor representation, critic representation and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

To train the agent, first specify the training options. For this example, use the following options:

- Run each training episode for at most `10000` episodes, with each episode lasting at most `maxsteps` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Verbose` and `Plots` options).
- Stop training when the agent receives an cumulative episode reward greater than `1700`.

For more information, see `rlTrainingOptions`.

```

maxepisodes = 1e4;
maxsteps = ceil(Tf/Ts);
trainingOpts = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','EpisodeReward', ...
    'StopTrainingValue',1700);

```

Train the agent using the `train` function. This is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

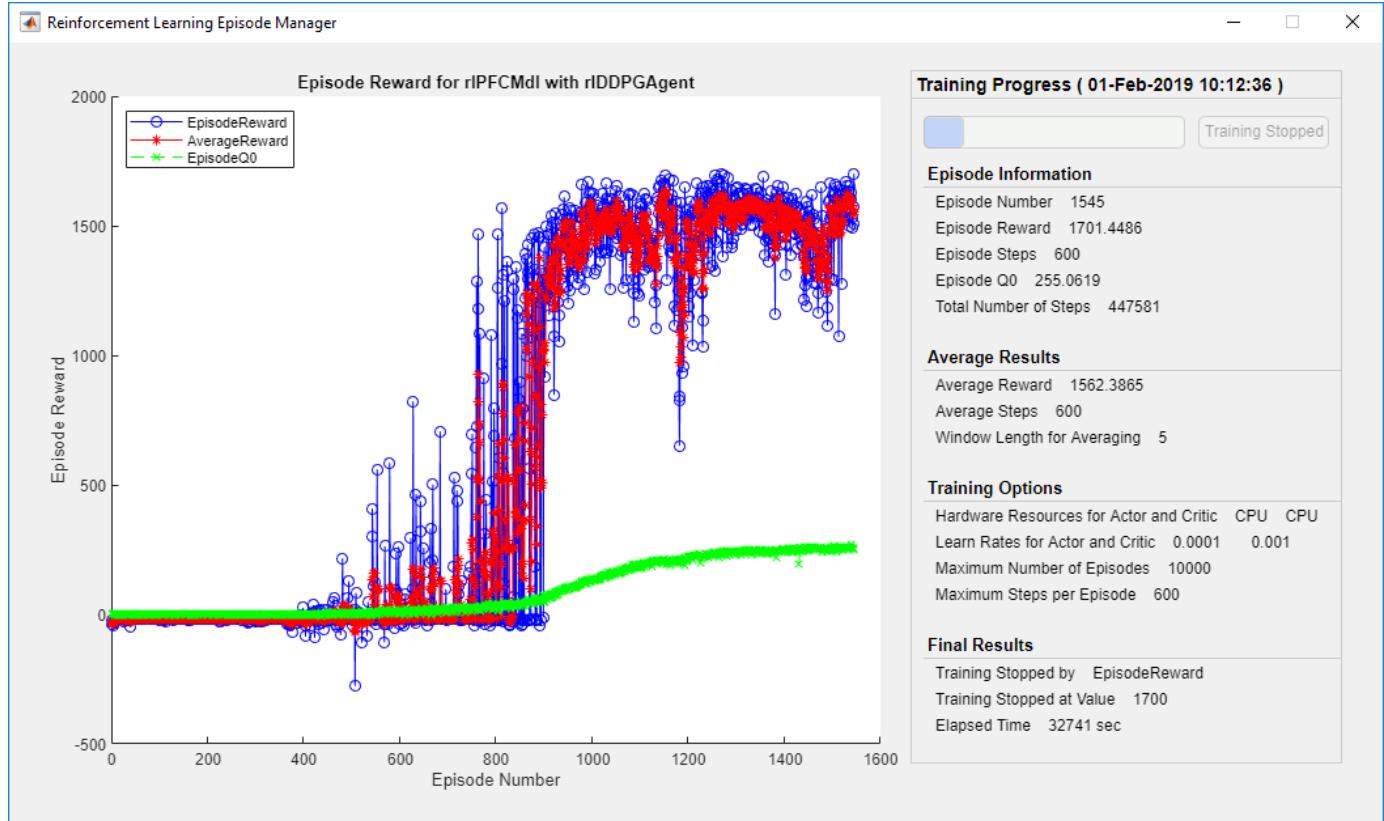
```

doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOpts);
else

```

```
% Load pretrained agent for the example.
load('SimulinkPFCDDPG.mat','agent')
end
```



Simulate DDPG Agent

To validate the performance of the trained agent, uncomment the following two lines and simulate it within the environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
% simOptions = rlSimulationOptions('MaxSteps',maxsteps);
% experience = sim(env,agent,simOptions);
```

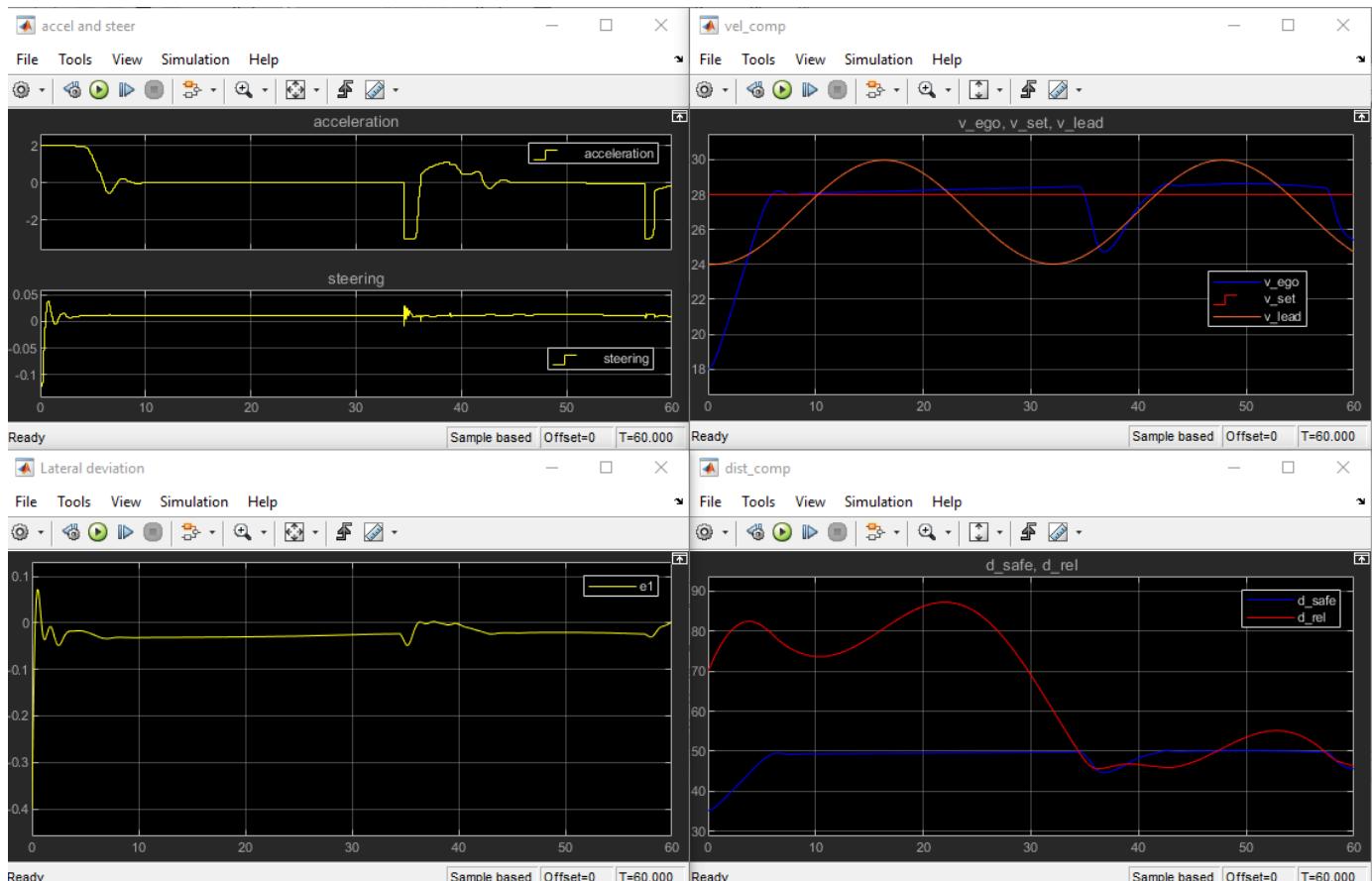
To demonstrate the trained agent using deterministic initial conditions, simulate the model in Simulink.

```
e1_initial = -0.4;
e2_initial = 0.1;
x0_lead = 80;
sim(mdl)
```

The following plots show the simulation results when lead car is 70 (m) ahead of ego car.

- In the first 35 seconds, relative distance is greater than safe distance (bottom right plot), thus ego car tracks set velocity (top right plot). To speed up and reach the set velocity, acceleration is mostly non-negative (top left plot).

- From 35 to 42 seconds, relative distance is mostly less than safe distance (bottom right plot), thus ego car tracks the minimum of lead velocity and set velocity. Since lead velocity is less than set velocity (top right plot), to track lead velocity, acceleration becomes non-zero (top left plot).
- From 42 to 58 seconds, ego car tracks set velocity (top right plot) and acceleration remains zero (top left plot).
- From 58 to 60 seconds, relative distance becomes less than safe distance (bottom right plot), thus ego car slows down and tracks lead velocity.
- The bottom left plot shows the lateral deviation. As shown in the plot, the lateral deviation is greatly decreased within one second. The lateral deviation remains less than 0.05 m.



Close Simulink model.

```
bdclose(mdl)
```

Local Function

```
function in = localResetFcn(in)
% reset
in = setVariable(in,'x0_lead',40+randi(60,1,1)); % random value for initial position of lead
in = setVariable(in,'el_initial', 0.5*(-1+2*rand)); % random value for lateral deviation
```

```
in = setVariable(in,'e2_initial', 0.1*(-1+2*rand)); % random value for relative yaw angle  
end
```

See Also

[train](#)

More About

- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Train DQN Agent for Lane Keeping Assist Using Parallel Computing

This example shows how to train a deep Q-learning network (DQN) agent for lane keeping assist (LKA) in Simulink® using parallel training. For an example that shows how to train the agent without using parallel training, see “Train DQN Agent for Lane Keeping Assist” on page 5-110.

For more information on DQN agents, see “Deep Q-Network Agents” on page 4-8. For an example that trains a DQN agent in MATLAB®, see “Train DQN Agent to Balance Cart-Pole System” on page 5-8.

DQN Parallel Training Overview

In a DQN agent, each worker generates new experiences from its copy of the agent and the environment. After every **N** steps, the worker sends experiences to the host agent. The host agent updates its parameters as follows.

- For asynchronous training, the host agent learns from received experiences without waiting for all workers to send experiences, and sends the updated parameters back to the worker that provided the experiences. Then, the worker continues to generate experiences from its environment using the updated parameters.
- For synchronous training, the host agent waits to receive experiences from all of the workers and learns from these experiences. The host then sends updated parameters to all the workers at the same time. Then, all workers continue to generate experiences using the updated parameters.

Simulink Model for Ego Car

The reinforcement learning environment for this example is a simple bicycle model for ego vehicle dynamics. The training goal is to keep the ego vehicle traveling along the centerline of the lanes by adjusting the front steering angle. This example uses the same vehicle model as “Train DQN Agent for Lane Keeping Assist” on page 5-110.

```
m = 1575; % total vehicle mass (kg)
Iz = 2875; % yaw moment of inertia (mNs^2)
lf = 1.2; % longitudinal distance from center of gravity to front tires (m)
lr = 1.6; % longitudinal distance from center of gravity to rear tires (m)
Cf = 19000; % cornering stiffness of front tires (N/rad)
Cr = 33000; % cornering stiffness of rear tires (N/rad)
Vx = 15; % longitudinal velocity (m/s)
```

Define the sample time **Ts** and simulation duration **T** in seconds.

```
Ts = 0.1;
T = 15;
```

The output of the LKA system is the front steering angle of the ego car. To simulate the physical steering limits of the ego car, constrain the steering angle to the range $[-0.5, 0.5]$ rad.

```
u_min = -0.5;
u_max = 0.5;
```

The curvature of the road is defined by a constant $0.001 (m^{-1})$. The initial value for the lateral deviation is 0.2 m and the initial value for the relative yaw angle is -0.1 rad.

```

rho = 0.001;
e1_initial = 0.2;
e2_initial = -0.1;

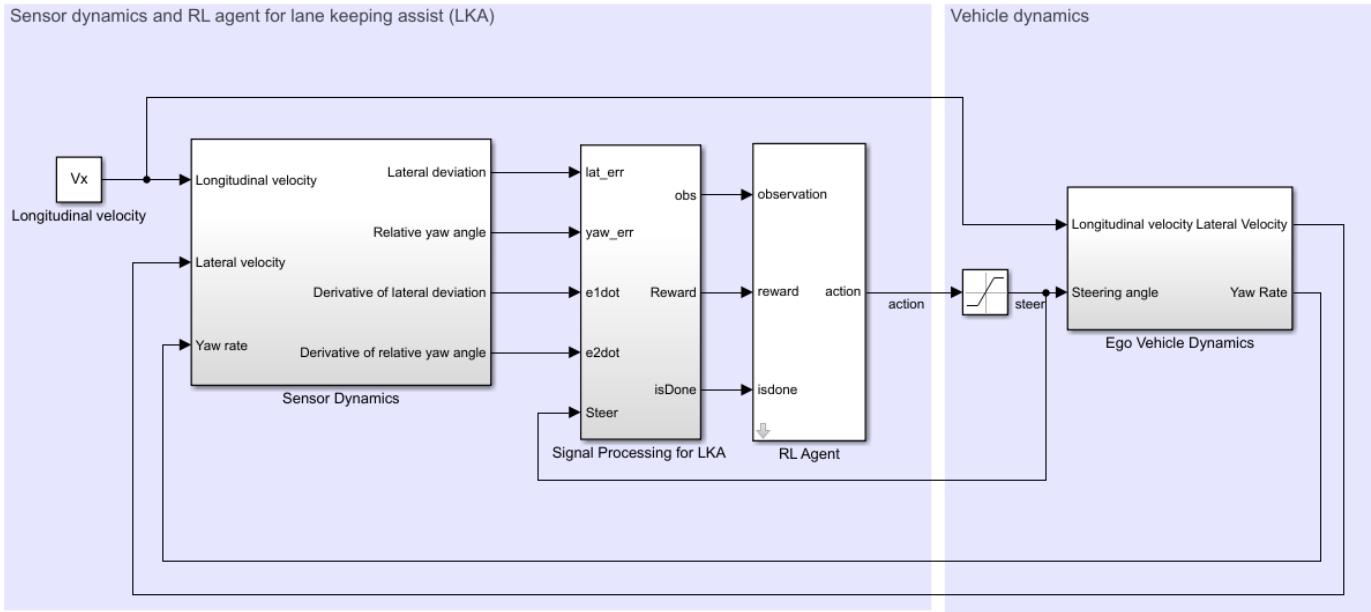
```

Open the model.

```

mdl = 'rlLKAMdl';
open_system(mdl)
agentblk = [mdl '/RL Agent'];

```



Copyright 2018 The MathWorks, Inc.

For this model:

- The steering-angle action signal from the agent to the environment is from -15 degrees to 15 degrees.
- The observations from the environment are the lateral deviation e_1 , relative yaw angle e_2 , their derivatives \dot{e}_1 and \dot{e}_2 , and their integrals $\int e_1$ and $\int e_2$.
- The simulation is terminated when the lateral deviation $|e_1| > 1$.
- The reward r_t , provided at every time step t , is

$$r_t = -(10e_1^2 + 5e_2^2 + 2u^2 + 5\dot{e}_1^2 + 5\dot{e}_2^2)$$

where u is the control input from the previous time step $t - 1$.

Create Environment Interface

Create a reinforcement learning environment interface for the ego vehicle.

Define the observation information.

```
observationInfo = rlNumericSpec([6 1], 'LowerLimit', -inf*ones(6,1), 'UpperLimit', inf*ones(6,1));
observationInfo.Name = 'observations';
observationInfo.Description = 'information on lateral deviation and relative yaw angle';
```

Define the action information.

```
actionInfo = rlFiniteSetSpec((-15:15)*pi/180);
actionInfo.Name = 'steering';
```

Create the environment interface.

```
env = rlSimulinkEnv(mdl, agentblk, observationInfo, actionInfo);
```

The interface has a discrete action space where the agent can apply one of 31 possible steering angles from -15 degrees to 15 degrees.

To define the initial condition for the lateral deviation and relative yaw angle, specify an environment reset function using an anonymous function handle. `localResetFcn`, which is defined at the end of this example, randomizes the initial lateral deviation and relative yaw angle.

```
env.ResetFcn = @(in)localResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DQN Agent

A DQN agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs (the state and action) and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
L = 24; % number of neurons
statePath = [
    imageInputLayer([6 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(L, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(L, 'Name', 'fc2')
    additionLayer(2, 'Name', 'add')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(L, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(1, 'Name', 'fc4')];

actionPath = [
    imageInputLayer([1 1 1], 'Normalization', 'none', 'Name', 'action')
    fullyConnectedLayer(L, 'Name', 'fc5')];

criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork, actionPath);
criticNetwork = connectLayers(criticNetwork, 'fc5', 'add/in2');
```

Specify options for the critic representation using `rlRepresentationOptions`.

```
criticOpts = rlRepresentationOptions('LearnRate', 1e-3, 'GradientThreshold', 1);
```

Create the critic representation using the specified deep neural network and options. You must also specify the action and observation info for the critic, which you obtain from the environment interface. For more information, see `rlQValueRepresentation`.

```
critic = rlQValueRepresentation(criticNetwork,observationInfo,actionInfo,'Observation',{ 'state'
```

To create the DQN agent, first specify the DQN agent options using `rlDQNAgentOptions`.

```
agentOpts = rlDQNAgentOptions(...
    'SampleTime',Ts, ...
    'UseDoubleDQN',true, ...
    'TargetSmoothFactor',1e-3, ...
    'DiscountFactor',0.99, ...
    'ExperienceBufferLength',1e6, ...
    'MiniBatchSize',64);
```

Then create the DQN agent using the specified critic representation and agent options. For more information, see `rlDQNAgent`.

```
agent = rlDQNAgent(critic,agentOpts);
```

Parallel Training Options

To train the agent, first specify the training options. For this example, use the following options.

- Run each training for at most 5000 episodes, with each episode lasting at most `ceil(T/Ts)` time steps.
- Display the training progress in the Episode Manager dialog box only (set the `Plots` and `Verbose` options accordingly).
- Stop training when the episode reward reaches -1.
- Save a copy of the agent for each episode where the cumulative reward is greater than -2.5.

For more information, see `rlTrainingOptions`.

```
maxepisodes = 5000;
maxsteps = ceil(T/Ts);
trainOpts = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Verbose',false, ...
    'Plots','training-progress',...
    'StopTrainingCriteria','EpisodeReward',...
    'StopTrainingValue', -1, ...
    'SaveAgentCriteria','EpisodeReward',...
    'SaveAgentValue',-2.5);
```

To train the agent in parallel, specify the following training options.

- Set the `UseParallel` option to `true`.
- Train agent in parallel asynchronously by setting the `ParallelizationOptions.Mode` option to "async".
- After every 30 steps, each worker sends experiences to the host.
- DQN agent requires workers to send "experiences" to the host.

```
trainOpts.UseParallel = true;
trainOpts.ParallelizationOptions.Mode = "async";
```

```
trainOpts.ParallelizationOptions.DataToSendFromWorkers = "experiences";
trainOpts.ParallelizationOptions.StepsUntilDataIsSent = 30;
trainOpts.ParallelizationOptions.WorkerRandomSeeds = -1;
```

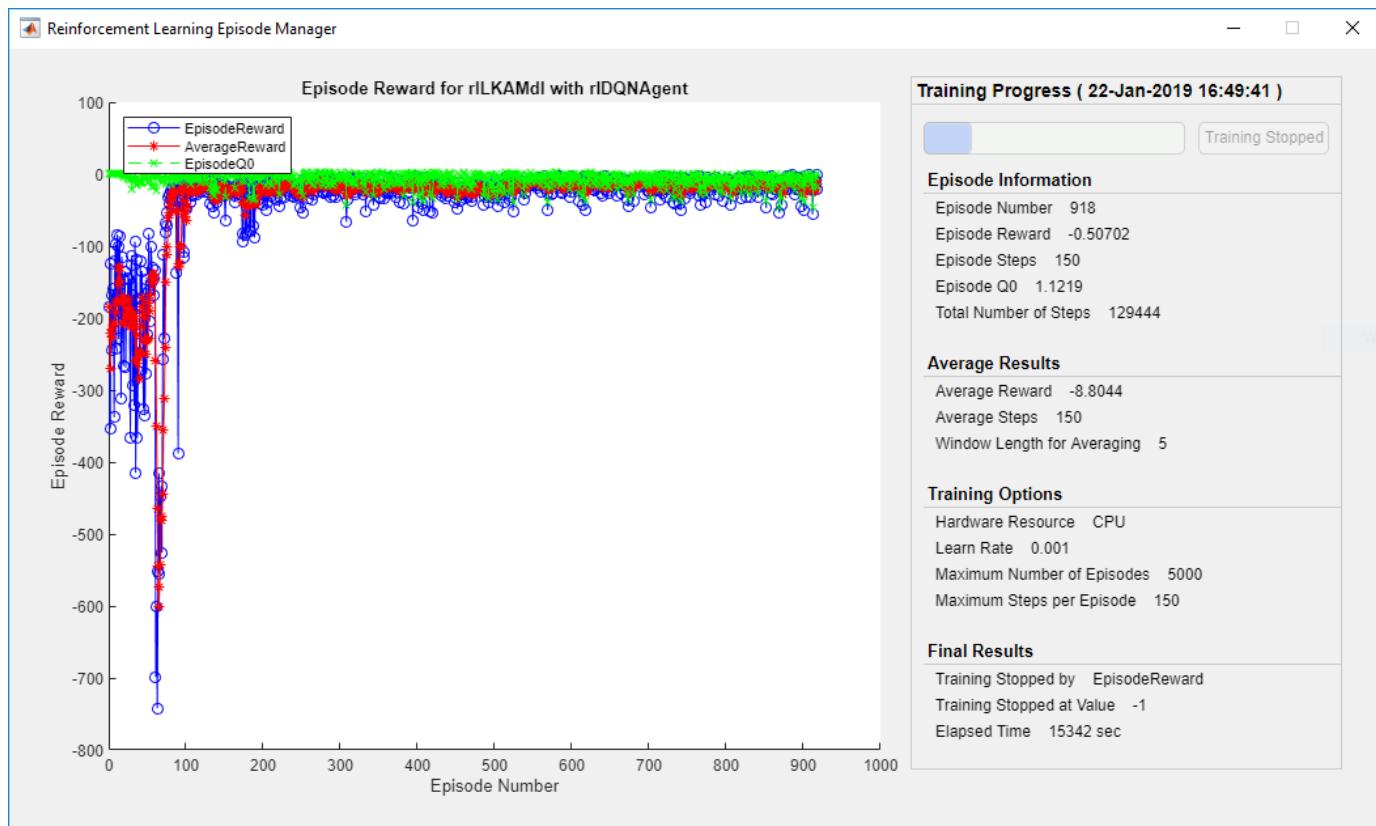
For more information, see [rlTrainingOptions](#).

Train Agent

Train the agent using the `train` function. Training the agent is a computationally intensive process that takes several minutes to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`. Due to randomness of the parallel training, you can expect different training results from the plot below. The plot shows the result of training with four workers.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example.
    load('SimulinkLKADQNParallel.mat','agent')
end
```



Simulate DQN Agent

To validate the performance of the trained agent, uncomment the following two lines and simulate the agent within the environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

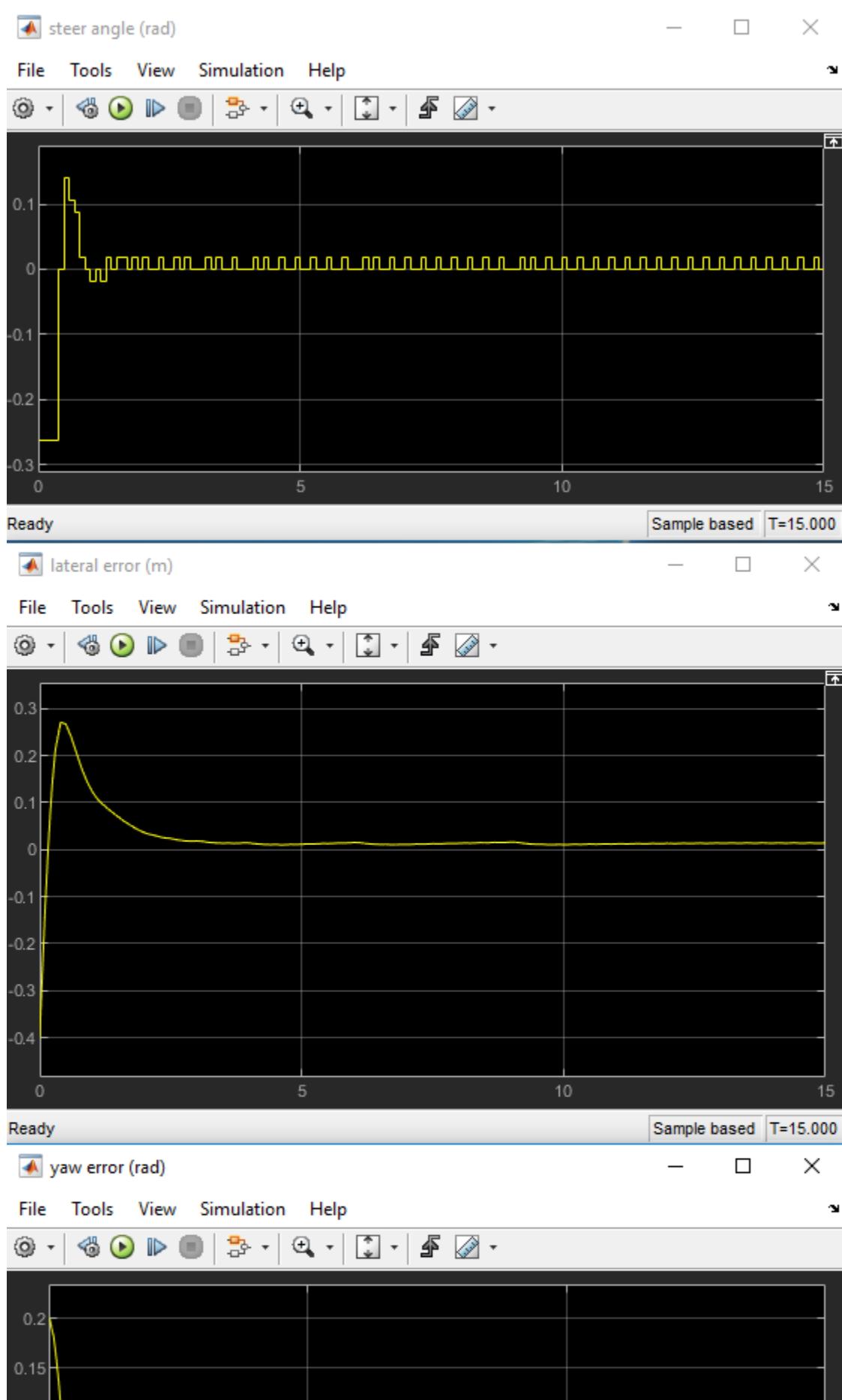
```
% simOptions = rlSimulationOptions('MaxSteps',maxsteps);  
% experience = sim(env,agent,simOptions);
```

To demonstrate the trained agent using deterministic initial conditions, simulate the model in Simulink.

```
e1_initial = -0.4;  
e2_initial = 0.2;  
sim(mdl)
```

As shown below, the lateral error (middle plot) and relative yaw angle (bottom plot) are both driven to zero. The vehicle starts from off centerline (-0.4 m) and nonzero yaw angle error (0.2 rad). The LKA enables the ego car to travel along the centerline after 2.5 seconds. The steering angle (top plot) shows that the controller reaches steady state after 2 seconds.

5 Train and Validate Agents



Local Function

```
function in = localResetFcn(in)
% reset
in = setVariable(in,'e1_initial', 0.5*(-1+2*rand)); % random value for lateral deviation
in = setVariable(in,'e2_initial', 0.1*(-1+2*rand)); % random value for relative yaw angle
end
```

See Also

[train](#)

More About

- “Train DQN Agent for Lane Keeping Assist” on page 5-110
- “Train Reinforcement Learning Agents” on page 5-2
- “Create Policy and Value Function Representations” on page 3-2

Train Biped Robot to Walk Using Reinforcement Learning Agents

This example shows how to train a biped robot, modeled in Simscape™ Multibody™, to walk using both a deep deterministic policy gradient (DDPG) agent and a Twin-Delayed deep deterministic policy gradient (TD3) agent and compares the performance of these trained agents.

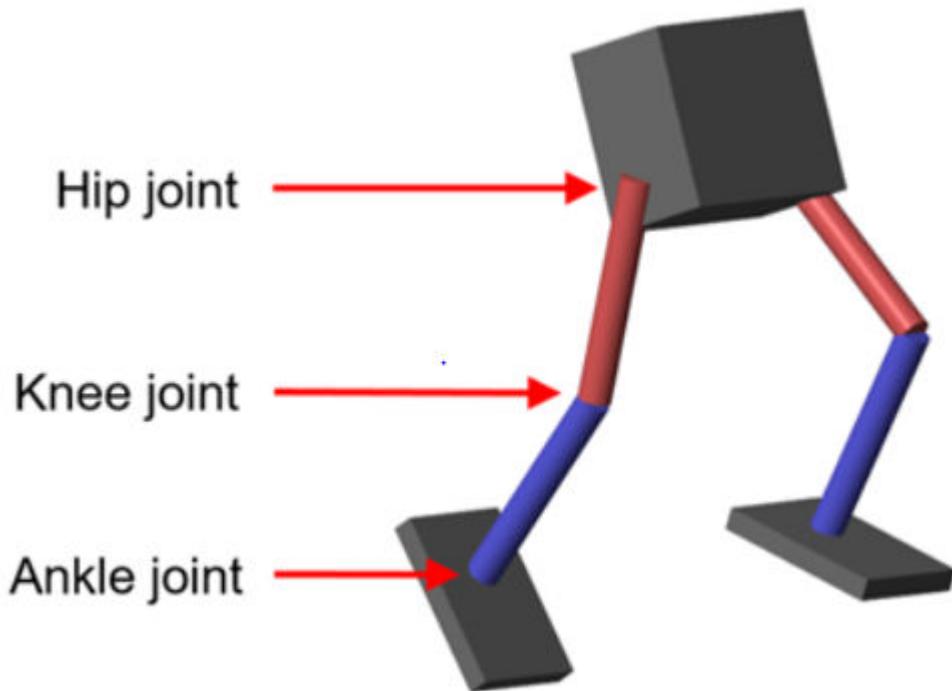
For more information on these agents, see “Deep Deterministic Policy Gradient Agents” on page 4-14 and “Twin-Delayed Deep Deterministic Policy Gradient Agents” on page 4-17.

For the purpose of comparison in this example, both agents are trained on the biped robot environment with same model parameters. The agents are also configured to have the following settings in common.

- Initial condition strategy of the biped robot
- Network structure of actor and critic, inspired by [2]
- Options for actor and critic representations
- Training options (sample time, discount factor, mini-batch size, experience buffer length, exploration noise)

Biped Robot Model

The reinforcement learning environment for this example is a biped robot. The training goal is to make the robot walk in a straight line using minimal control effort.



Load parameters of the model to the MATLAB® workspace.

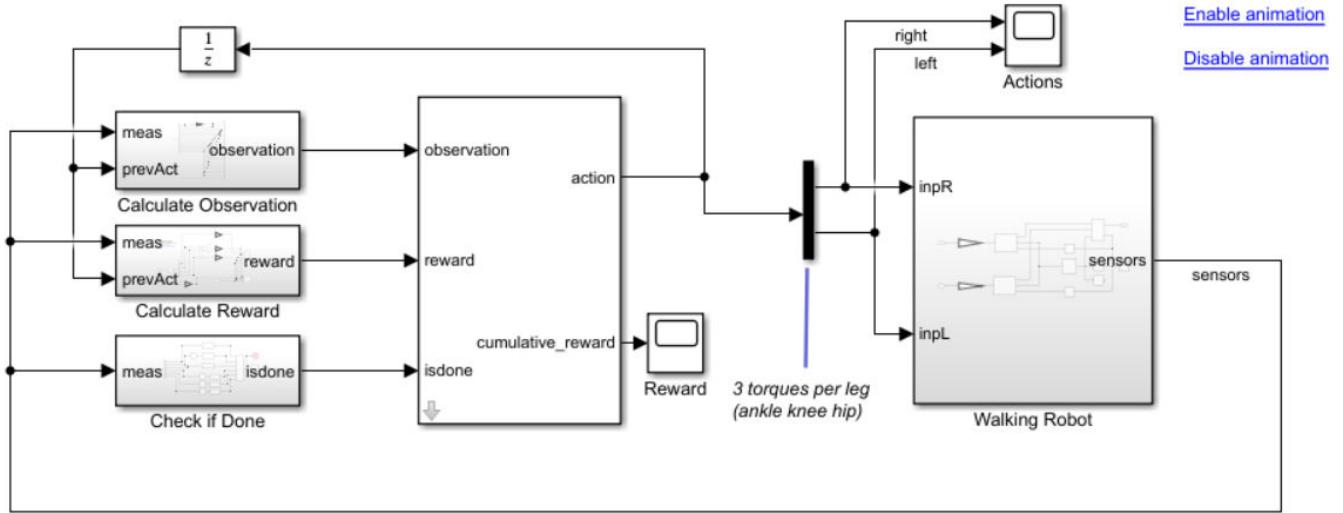
```
robotParametersRL
```

Open the Simulink model.

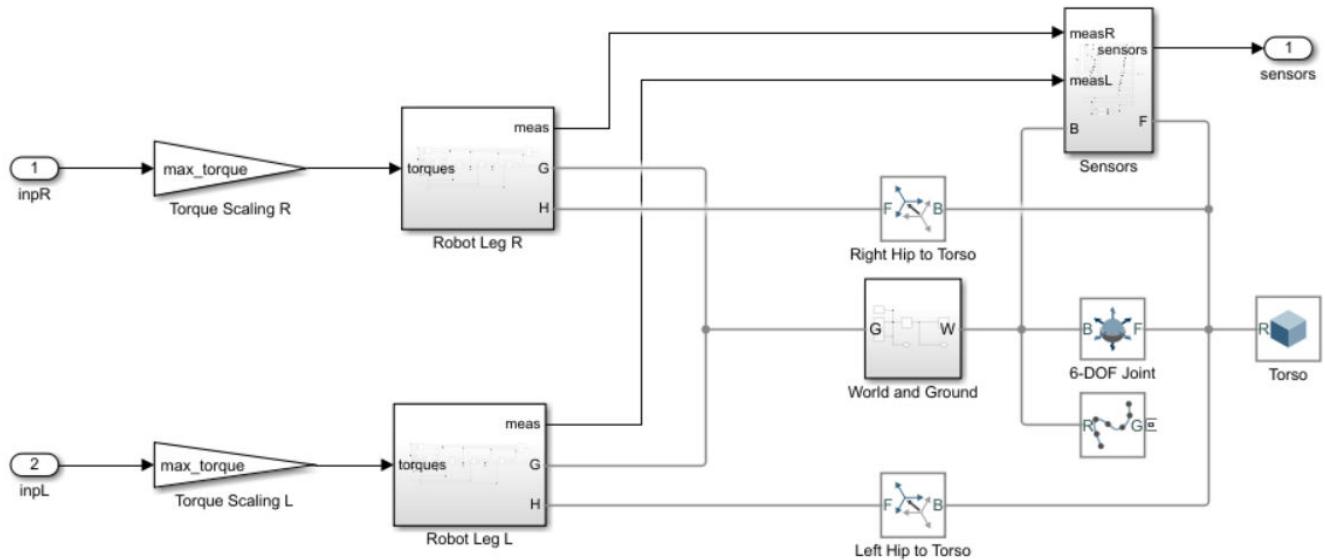
```
mdl = 'rlWalkingBipedRobot';
open_system(mdl)
```

Walking Robot: Reinforcement Learning (2D)

Copyright 2019 The MathWorks, Inc.



The robot is modeled using Simscape Multibody.



For this model:

- The neutral 0 radian position is with both legs straight and the ankles flat.
- The foot contact is modeled using the Spatial Contact Force block from Simscape Multibody.

- The agent can control 3 individual joints, the ankle, knee, and hip, on both legs of the robot by applying torque signals from -3 to 3 Nm. The actual computed action signals are normalized between -1 and 1.

The environment provides 29 observations to the agent. The observations are:

- Y (lateral) and Z (vertical) translations of the torso center of mass. The translation in the Z direction is normalized to a similar range as the other observations.
- X (forward), Y (lateral), and Z (vertical) translation velocities.
- Yaw, pitch, and roll angles of the torso.
- Yaw, pitch, and roll angular velocities of the torso.
- Angular positions and velocities of the 3 joints (ankle, knee, hip) on both legs.
- Action values from the previous time step.

The episode terminates if either of the following conditions occur.

- The robot torso center of mass is less than 0.1 m in Z direction (fallen) or more than 1 m in Y direction (lateral motion).
- The absolute value of either the roll, pitch, or yaw is greater than 0.7854 radians.

The reward function r_t , which is provided at every time step, is inspired by [1]. This reward function encourages the agent to move forward by providing a positive reward for positive forward velocity. It also encourages the agent to avoid episode termination by providing a constant reward ($25 \frac{T_s}{T_f}$) at every time step. The other terms in the reward function are penalties for substantial changes in lateral and vertical translations, and for the use of excess control effort.

$$r_t = v_x - 3y^2 - 50\hat{z}^2 + 25 \frac{T_s}{T_f} - 0.02 \sum_i u_{t-1}^i$$

Here:

- v_x is the translation velocity in X direction (forward toward goal) of the robot.
- y is the lateral translation displacement of the robot from the target straight line trajectory.
- \hat{z} is the normalized vertical translation displacement of the robot center of mass.
- u_{t-1}^i is the torque from joint i from the previous time step.
- T_s is the sample time of the environment.
- T_f is the final simulation time of the environment.

Create Environment Interface

Create the observation specification.

```
numObs = 29;
obsInfo = rlNumericSpec([numObs 1]);
obsInfo.Name = 'observations';
```

Create the action specification.

```
numAct = 6;
actInfo = rlNumericSpec([numAct 1], 'LowerLimit', -1, 'UpperLimit', 1);
actInfo.Name = 'foot_torque';
```

Create the environment interface for the walking robot model.

```
blk = [mdl, '/RL Agent'];
env = rlSimulinkEnv(mdl,blk,obsInfo,actInfo);
env.ResetFcn = @(in) walkerResetFcn(in,upper_leg_length/100,lower_leg_length/100,h/100);
```

Select and Create Agent for Training

This example provides the option to train the robot either using either a DDPG or TD3 agent. To simulate the robot with the agent of your choice, set the `AgentSelection` flag accordingly.

```
AgentSelection = 'TD3';
switch AgentSelection
    case 'DDPG'
        agent = createDDPGAgent(numObs,obsInfo,numAct,actInfo,Ts);
    case 'TD3'
        agent = createTD3Agent(numObs,obsInfo,numAct,actInfo,Ts);
    otherwise
        disp('Enter DDPG or TD3 for AgentSelection')
end
```

The `createDDPGAgent` and `createTD3Agent` helper functions perform the following actions.

- Create actor and critic networks.
- Specify options for actor and critic representations.
- Create actor and critic representations using created networks and specified options.
- Configure agent specific options.
- Create agent.

DDPG Agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. A DDPG agent decides which action to take given observations using an actor representation. The actor and critic networks for this example are inspired by [2].

For details on the creating the DDPG agent, see the `createDDPGAgent` helper function. For information on configuring DDPG agent options, see `rlDDPGAgentOptions`.

For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2. For an example that creates neural networks for DDPG agents, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

TD3 Agent

TD3 agent approximates the long-term reward given observations and actions using 2 critic value function representations. A TD3 agent decides which action to take given observations using an actor representation. The structure of the actor and critic networks used for this agent are the same as the ones used for DDPG agent.

A DDPG agent can overestimate the Q value. Since this Q value is then used to update the policy (actor) of the agent, the resultant policy can be suboptimal and accumulating training errors can lead to divergent behavior. The TD3 algorithm is an extension of DDPG with improvements that make it more robust by preventing overestimation of Q values [3].

- Two critic networks — TD3 agents learn two critic networks independently and use the minimum value function estimate to update the actor (policy). Doing so prevents accumulation of error in subsequent steps and overestimation of Q values.
- Addition of target policy noise — Clipped noise is added to target actions to smooth out Q function values over similar actions. Doing so prevents learning an incorrect sharp peak of noisy value estimate.
- Delayed policy and target updates — For a TD3 agent it is recommended to delay the actor network update, as it allows more time for the Q function to reduce error (get closer to the required target) before updating the policy. Doing so prevents variance in values estimates and results in a more high quality policy update.

For details on the creating the TD3 agent, see the `createTD3Agent` helper function. For information on configuring TD3 agent options, see `rlTD3AgentOptions`.

Specify Training Options and Train Agent

For this example, the training options for the DDPG and TD3 agents are the same. These options are based on the following requirements.

- Run each training session for 2000 episodes with each episode lasting at most `maxSteps` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option).
- Terminate the training only when it reaches the maximum number of episodes (`maxEpisodes`). Doing so allows the comparison of the learning curves for multiple agents over the entire training session.

For more information and additional options, see `rlTrainingOptions`.

```
maxEpisodes = 2000;
maxSteps = floor(Tf/Ts);
trainOpts = rlTrainingOptions(...
    'MaxEpisodes',maxEpisodes, ...
    'MaxStepsPerEpisode',maxSteps, ...
    'ScoreAveragingWindowLength',250, ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','EpisodeCount', ...
    'StopTrainingValue',maxEpisodes, ...
    'SaveAgentCriteria','EpisodeCount', ...
    'SaveAgentValue',maxEpisodes);
```

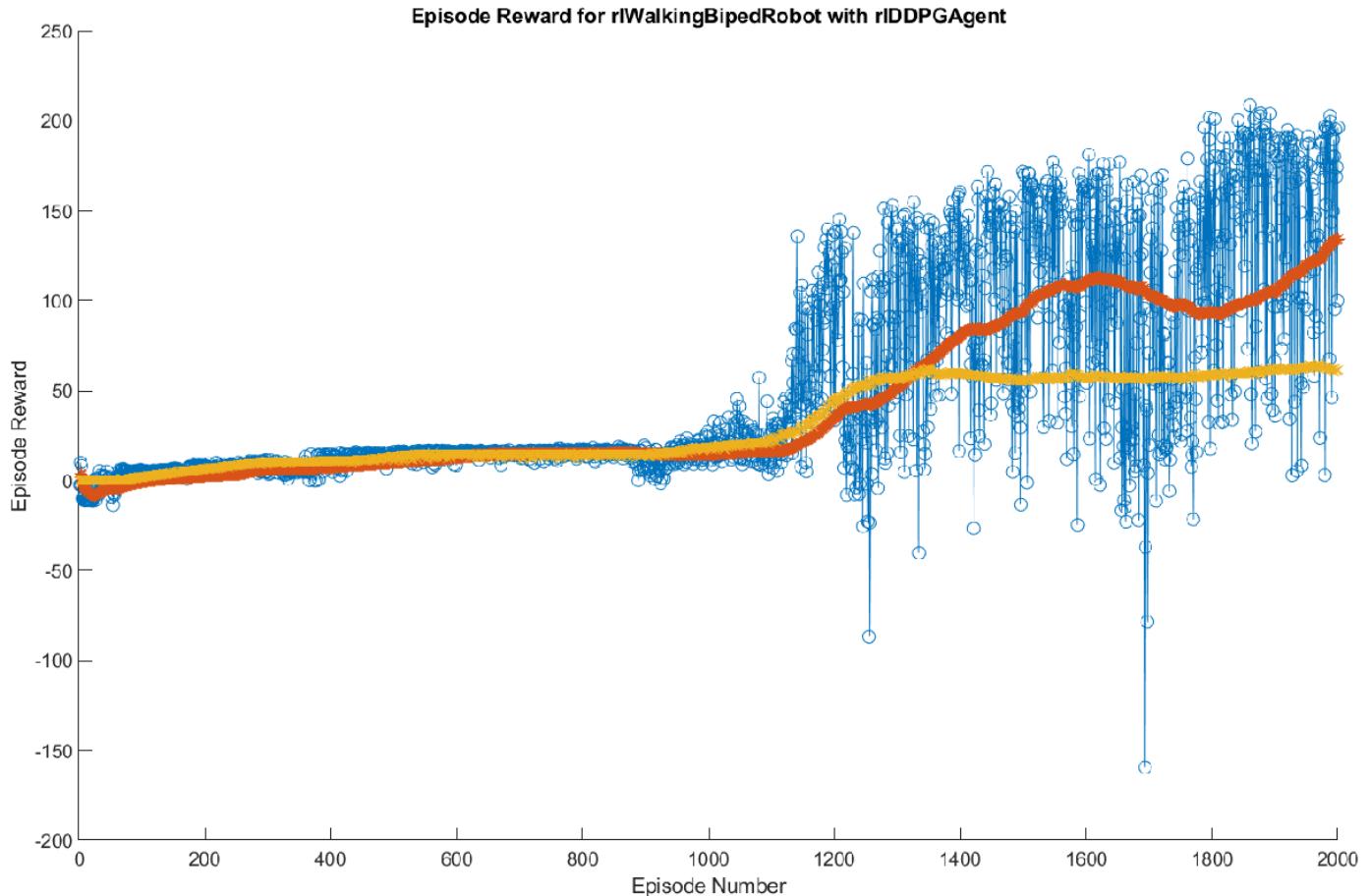
Specify the following training options to train the agent in parallel training mode. If you do not have Parallel Computing Toolbox™ software installed, set `UseParallel` to `false`.

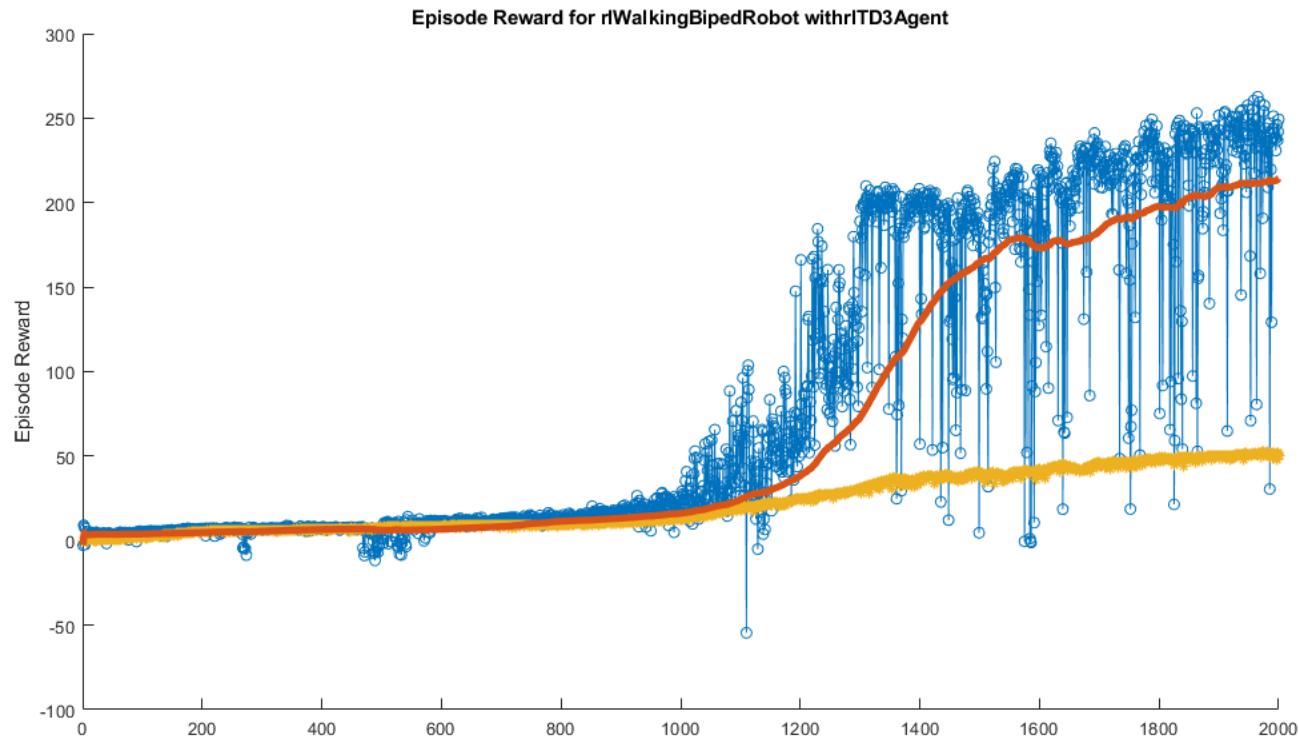
- Set the `UseParallel` option to `true`.
- Train the agent in parallel asynchronously.
- After every 32 steps, have each worker send experiences to the host. DDPG and TD3 agents require workers to send experiences to the host.

```
trainOpts.UseParallel = true;
trainOpts.ParallelizationOptions.Mode = 'async';
trainOpts.ParallelizationOptions.StepsUntilDataIsSent = 32;
trainOpts.ParallelizationOptions.DataToSendFromWorkers = 'Experiences';
```

Train the agent using the `train` function. This process is computationally intensive and takes several hours to complete for each agent. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`. Due to randomness in the parallel training, you can expect different training results from the plots below. The pretrained agents were trained in parallel using four workers.

```
doTraining = false;
if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the selected agent.
    if strcmp(AgentSelection,'DDPG')
        load('rlWalkingBipedRobotDDPG.mat','agent')
    else
        load('rlWalkingBipedRobotTD3.mat','agent')
    end
end
```





For the preceding example training curves, the average time per training step for the DDPG and TD3 agents are 0.11 and 0.12 seconds, respectively. The TD3 agent takes more training time per step because it updates two critic networks compared to the single critic used for DDPG.

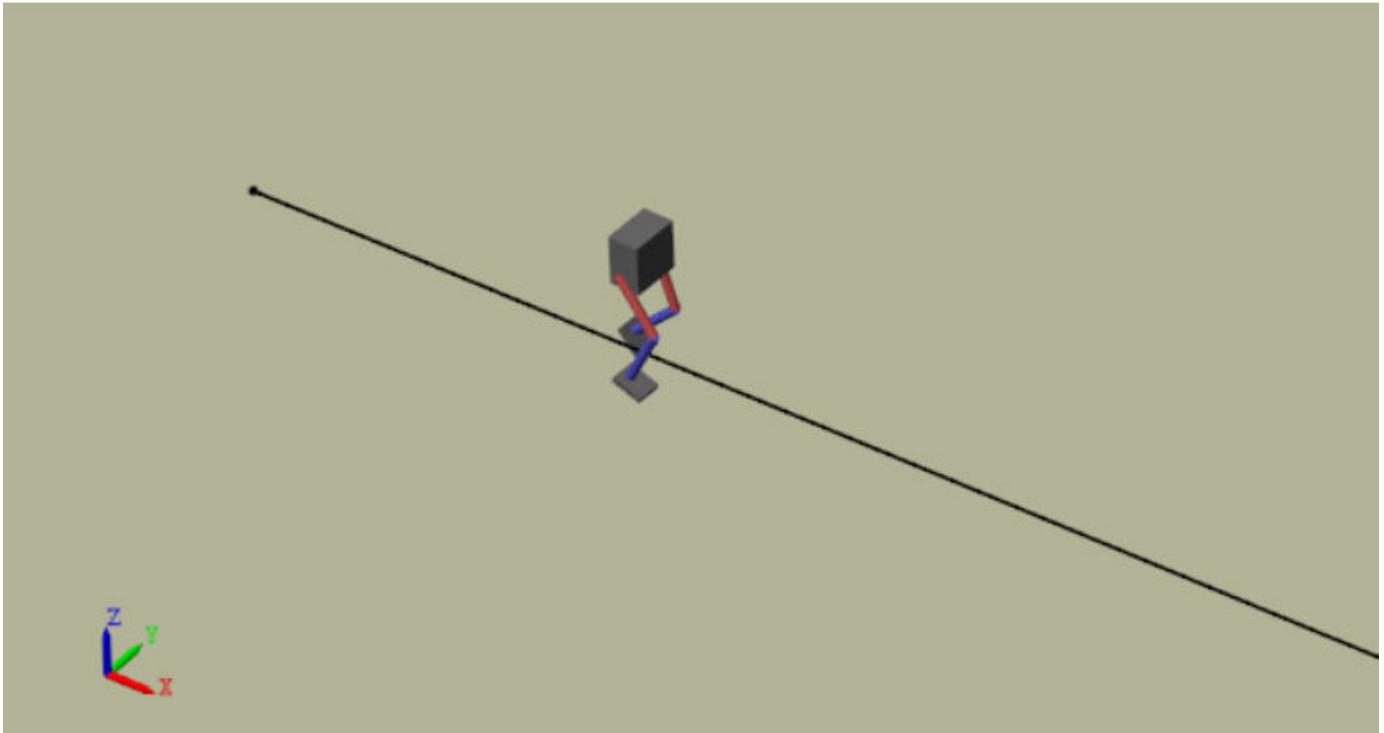
Simulate Trained Agents

Fix the random generator seed for reproducibility.

```
rng(0)
```

To validate the performance of the trained agent, simulate it within the biped robot environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',maxSteps);
experience = sim(env,agent,simOptions);
```

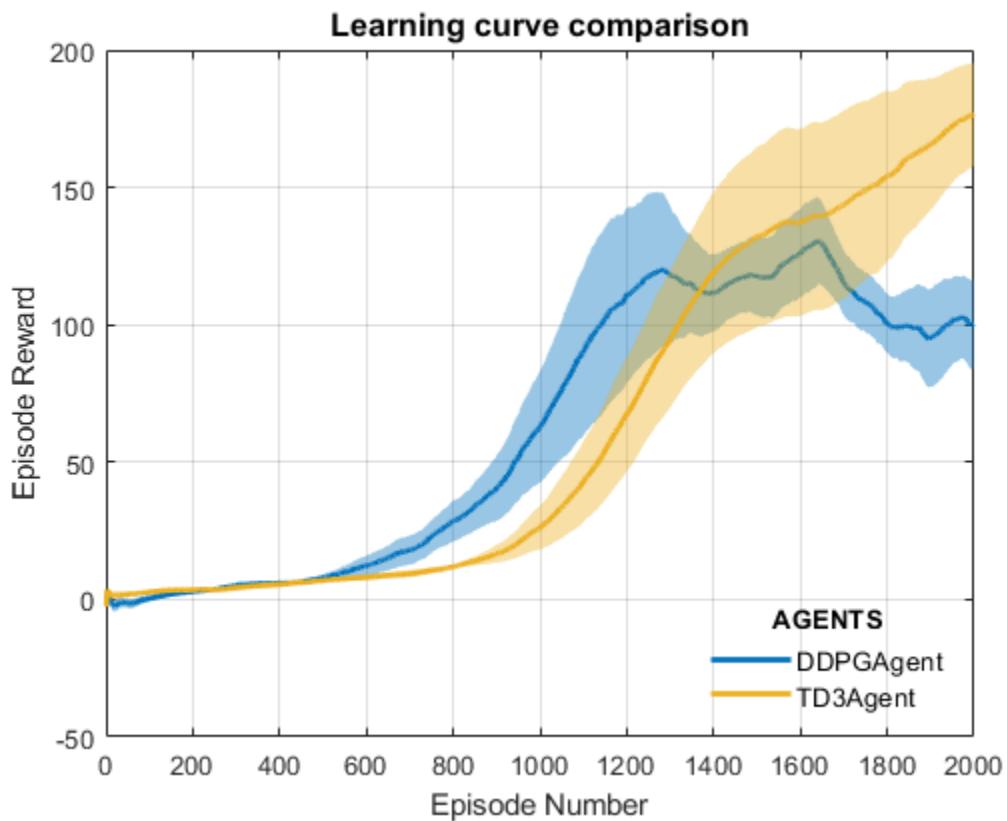


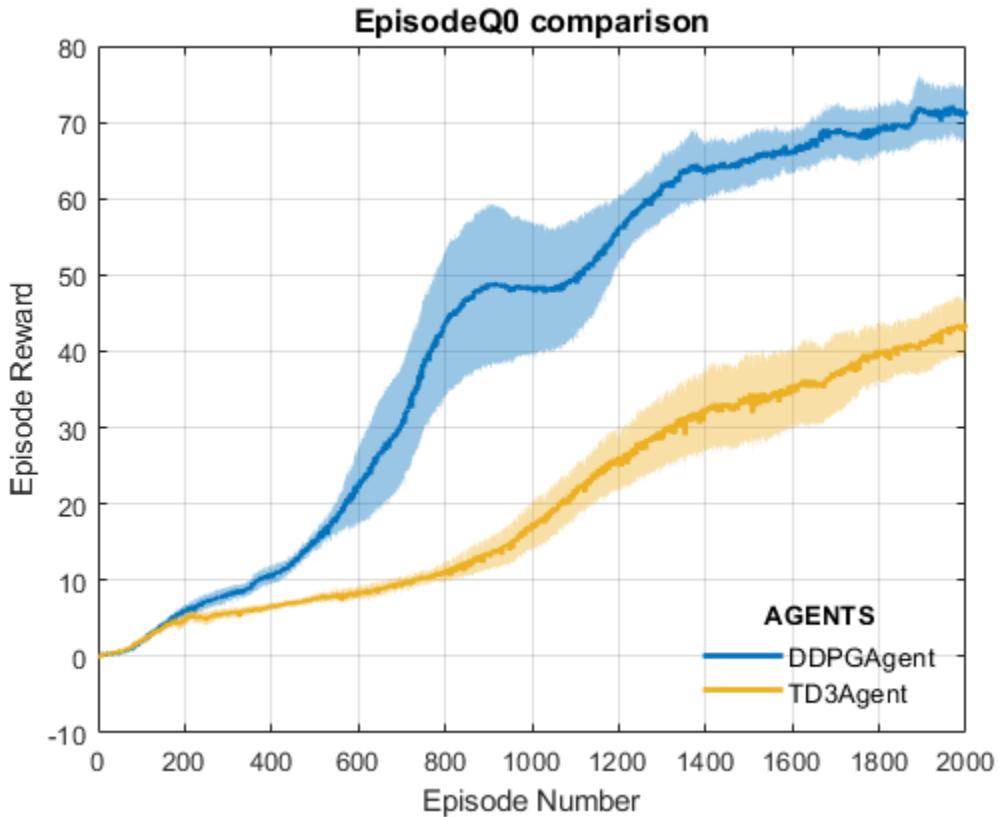
Compare Agent Performance

For the following agent comparison, each agent was trained five times using a different random seed each time. Due to the random exploration noise and the randomness in the parallel training, the learning curve for each run is different. Since the training of agents for multiple runs takes several days to complete, this comparison uses pretrained agents.

For the DDPG and TD3 agents, plot the average and standard deviation of the episode reward (top plot) and the episode Q0 value (bottom plot). The episode Q0 value is the critic estimate of the discounted long-term reward at the start of each episode given the initial observation of the environment. For a well-designed critic, the episode Q0 value approaches the true discounted long-term reward.

```
comparePerformance('DDPGAgent', 'TD3Agent')
```





Based on the **Learning curve comparison** plot:

- 1 The DDPG appears to pick up learning faster (around episode number 600 on an average) but hits a local minimum. TD3 starts slower but eventually achieves higher rewards than DDPG as it avoids overestimation of Q values.
- 2 The TD3 Agent shows a steady improvement in its learning curve, which suggests improved stability when compared to the DDPG agent.

Based on the **Episode Q0 comparison** plot:

- 1 For the TD3 agent, the critic estimate of the discounted long term reward (for 2000 episodes) is lower compared to the DDPG agent. This difference is because the TD3 algorithm takes a conservative approach in updating its targets by using minimum of two Q functions. This behavior is further enhanced because of delayed updates to the targets.
- 2 Although the TD3 estimate for these 2000 episodes is low, the TD3 agent shows a steady increase in the episode Q0 values unlike DDPG agent.

In this example the training was stopped at 2000 episodes. For a larger training period, the TD3 agent with its steady increase in estimates shows the potential to converge to the true discounted long-term reward.

References

- [1] Heess, Nicolas, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, et al. 'Emergence of Locomotion Behaviours in Rich Environments'. *ArXiv:1707.02286 [Cs]*, 10 July 2017. <https://arxiv.org/abs/1707.02286>.

[2] Lillicrap, Timothy P., Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. ‘Continuous Control with Deep Reinforcement Learning’. *ArXiv:1509.02971 [Cs, Stat]*, 5 July 2019. <https://arxiv.org/abs/1509.02971>.

[3] Fujimoto, Scott, Herke van Hoof, and David Meger. ‘Addressing Function Approximation Error in Actor-Critic Methods’. *ArXiv:1802.09477 [Cs, Stat]*, 22 October 2018. <https://arxiv.org/abs/1802.09477>.

See Also

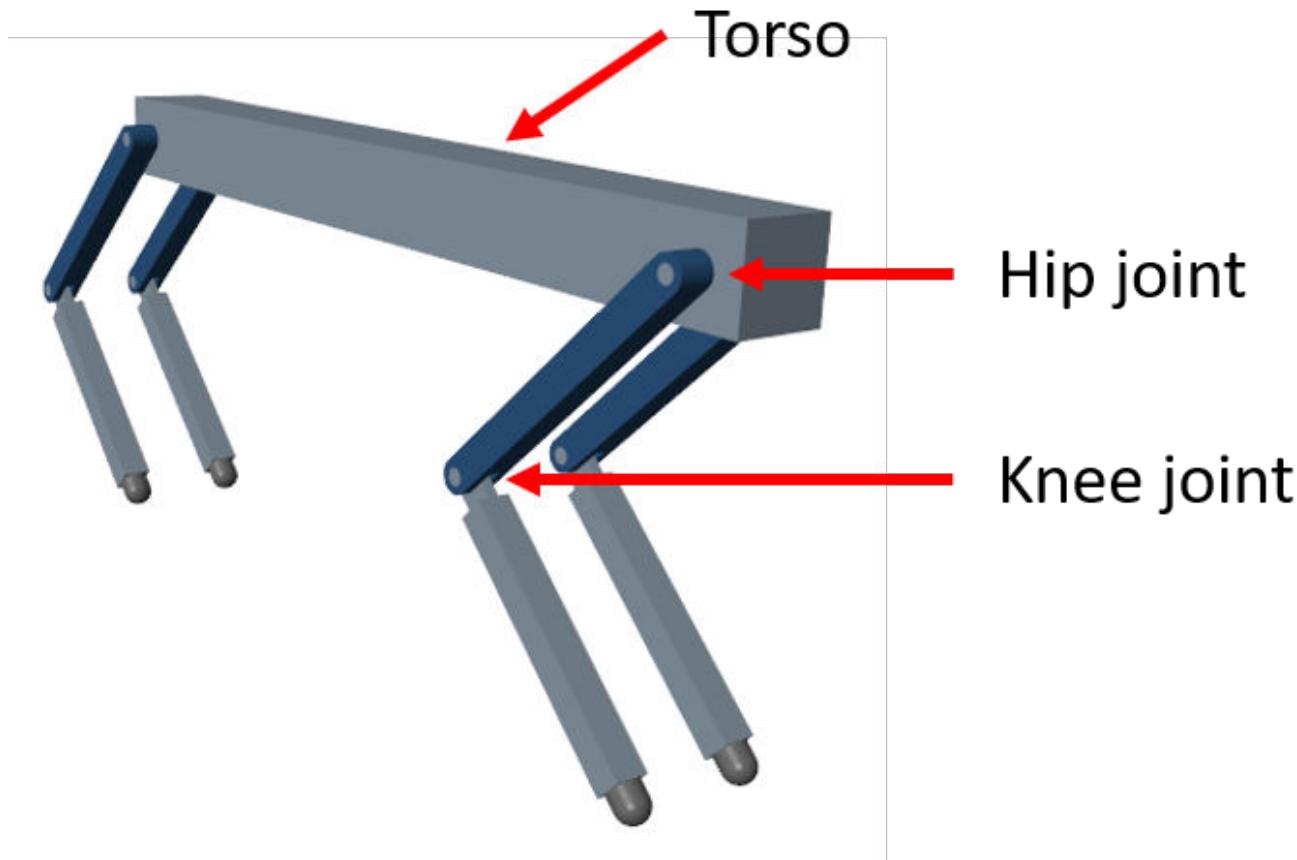
train

More About

- “Reinforcement Learning Agents” on page 4-2
- “Train Reinforcement Learning Agents” on page 5-2
- “Define Reward Signals” on page 2-7

Quadruped Robot Locomotion Using DDPG Agent

This example shows how to train a quadruped robot, modeled using Simscape Multibody, to walk using a deep deterministic policy gradient (DDPG) agent. For more information on DDPG agents, see “Twin-Delayed Deep Deterministic Policy Gradient Agents” on page 4-17.



Load necessary parameters to the base workspace in MATLAB®.

```
initializeRobotParameters
```

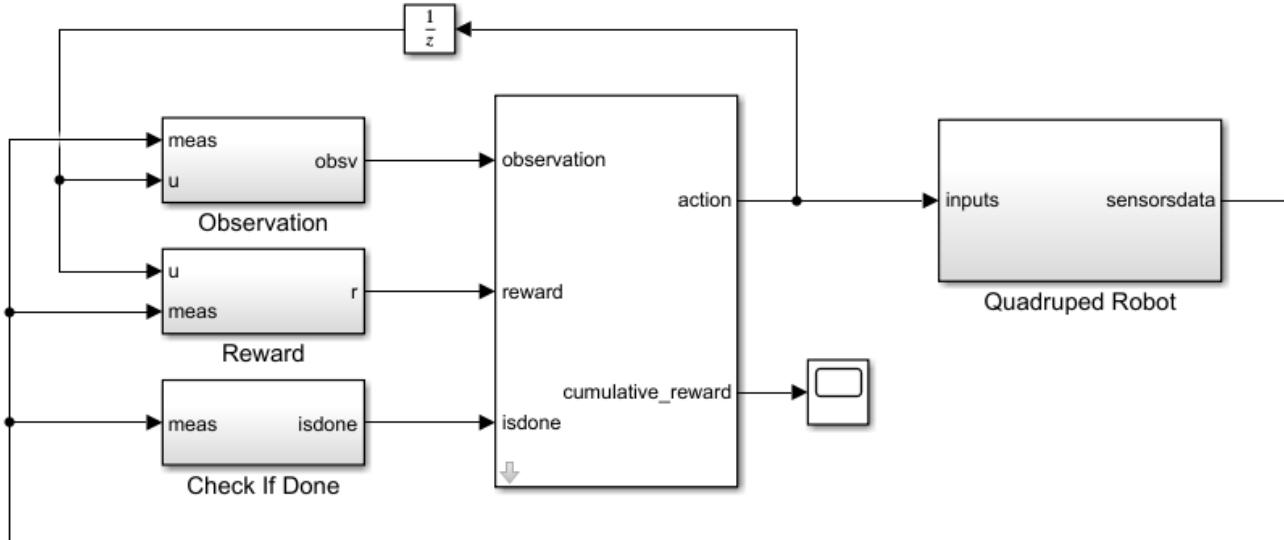
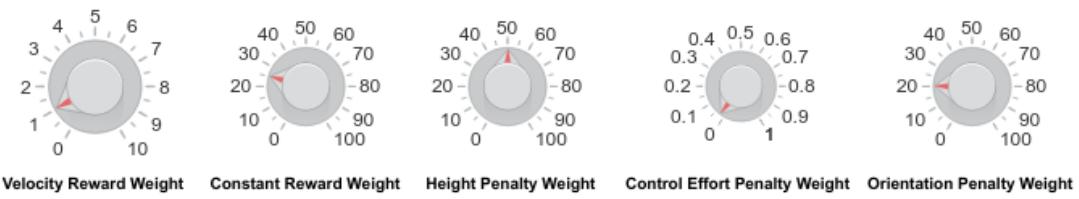
Quadruped Robot Model

The environment for this example is a quadruped robot, and the training goal is to make the robot walk in a straight line using minimal control effort.

The robot is modeled using Simscape Multibody and the Simscape Multibody Contact Forces Library. The main structural components are four legs and a torso as shown in the figure. The legs are connected to the torso through revolute joints. Action values provided by the RL Agent block are scaled and converted into joint torque values. These joint torque values are used by the revolute joints to compute motion.

Open the model.

```
mdl = 'rlQuadrupedRobot';
open_system(mdl)
```



Quadruped Walking Robot Example

Copyright 2019 The MathWorks, Inc.

Observations

The robot environment provides 44 observations to the agent, each normalized between -1 and 1. These observations are:

- Y (vertical) and Y (lateral) position of the torso center of mass
- Quaternion representing the orientation of the torso
- X (forward), Y (vertical), and Z (lateral) velocities of the torso at the center of mass
- Roll, pitch, and yaw rates of the torso
- Angular positions and velocities of the hip and knee joints for each leg
- Normal and friction force due to ground contact for each leg
- Action values (torque for each joint) from the previous time step

For all four legs, the initial values for the hip and knee joint angles are set to -0.8234 and 1.6468 radians, respectively. The neutral positions of the joints are at 0 radians. This occurs when the legs are stretched to their maximum and are aligned perpendicularly to the ground.

Actions

The agent generates eight actions normalized between -1 and 1. After multiplying with a scaling factor, these correspond to the eight joint torque signals for the revolute joints. The overall joint torque bounds are +/- 10 Nm for each joint.

Reward

The following reward function is provided to the agent performance at each time step during training. This reward function encourages the agent to move forward by providing a positive reward for positive forward velocity. It also encourages the agent to avoid early termination by providing a constant reward ($25T_s/T_f$) at each time step. The remaining terms in the reward function are penalties that discourage unwanted states, such as large deviations from the desired height and orientation or use of excessive joint torques.

$$r_t = v_x + 25 \frac{T_s}{T_f} - 50\hat{y}^2 - 20\theta^2 - 0.02 \sum_i u_{t-1}^i$$

where

- v_x is the velocity of the torso's center of mass in the x-direction.
- T_s and T_f are the sample time and final simulation time of the environment, respectively.
- \hat{y} is the scaled height error of the torso's center of mass from the desired height of 0.75m.
- θ is the pitch angle of the torso.
- u_{t-1}^i is the action value for joint i from the previous time step.

Episode termination

During training or simulation, the episode terminates if

- The height of the torso center of mass from the ground is below 0.5 m (fallen)
- The head or tail of the torso is below the ground
- Any knee joint is below the ground
- Roll, pitch, or yaw angles are outside bounds (+/- 0.1745, 0.1745 and 0.3491 radians, respectively).

Create Environment Interface

Specify the parameters for the observation set.

```
numObs = 44;
obsInfo = rlNumericSpec([numObs 1]);
obsInfo.Name = 'observations';
```

Specify the parameters for the action set.

```
numAct = 8;
actInfo = rlNumericSpec([numAct 1], 'LowerLimit', -1, 'UpperLimit', 1);
actInfo.Name = 'torque';
```

Create the environment using the reinforcement learning model.

```
blk = [mdl, '/RL Agent'];
env = rlSimulinkEnv(mdl,blk,obsInfo,actInfo);
```

During training, the reset function introduces random deviations into the initial joint angles and angular velocities.

```
env.ResetFcn = @quadrupedResetFcn;
```

Create DDPG agent

The DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. The agent also decides which action to take given the observations, using an actor representation. The actor and critic networks for this example are inspired by [2].

For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2. For an example that creates neural networks for DDPG agents, see “Train DDPG Agent to Control Double Integrator System” on page 5-31.

Create the networks in the MATLAB workspace using the `createNetworks` helper function.

```
createNetworks
```

You can also create your actor and critic networks interactively using the Deep Network Designer app.

View the critic network configuration.

```
plot(criticNetwork)
```

Specify the agent options using `rlDDPGAgentOptions`.

```
agentOptions = rlDDPGAgentOptions;
agentOptions.SampleTime = Ts;
agentOptions.DiscountFactor = 0.99;
agentOptions.MiniBatchSize = 250;
agentOptions.ExperienceBufferLength = 1e6;
agentOptions.TargetSmoothFactor = 1e-3;
agentOptions.NoiseOptions.MeanAttractionConstant = 0.15;
agentOptions.NoiseOptions.Variance = 0.1;
```

Create the `rlDDPGAgent` object for the agent.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Specify Training Options

To train the agent, first specify the following training options:

- Run each training episode for at most 10000 episodes, with each episode lasting at most `maxSteps` time steps.
- Display the training progress in the Episode Manager dialog box (set the `Plots` option) and disable the command line display (set the `Verbose` option).
- Stop training when the agent receives an average cumulative reward greater than 190 over 250 consecutive episodes.
- Save a copy of the agent for each episode where the cumulative reward is greater than 200.

```

maxEpisodes = 10000;
maxSteps = floor(Tf/Ts);
trainOpts = rlTrainingOptions(...
    'MaxEpisodes',maxEpisodes, ...
    'MaxStepsPerEpisode',maxSteps, ...
    'ScoreAveragingWindowLength',250, ...
    'Verbose',true, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','AverageReward', ...
    'StopTrainingValue',190, ...
    'SaveAgentCriteria','EpisodeReward', ...
    'SaveAgentValue',200);

```

Specify the following training options to train the agent in parallel training mode. If you do not have Parallel Computing Toolbox™ software installed, set `UseParallel` to `false`.

- Set the `UseParallel` option to `true`.
- Train the agent in parallel asynchronously.
- After every 32 steps, each worker sends experiences to the host.
- DDPG agents require workers to send '`Experiences`' to the host.

```

trainOpts.UseParallel = true;
trainOpts.ParallelizationOptions.Mode = 'async';
trainOpts.ParallelizationOptions.StepsUntilDataIsSent = 32;
trainOpts.ParallelizationOptions.DataToSendFromWorkers = 'Experiences';

```

Train Agent

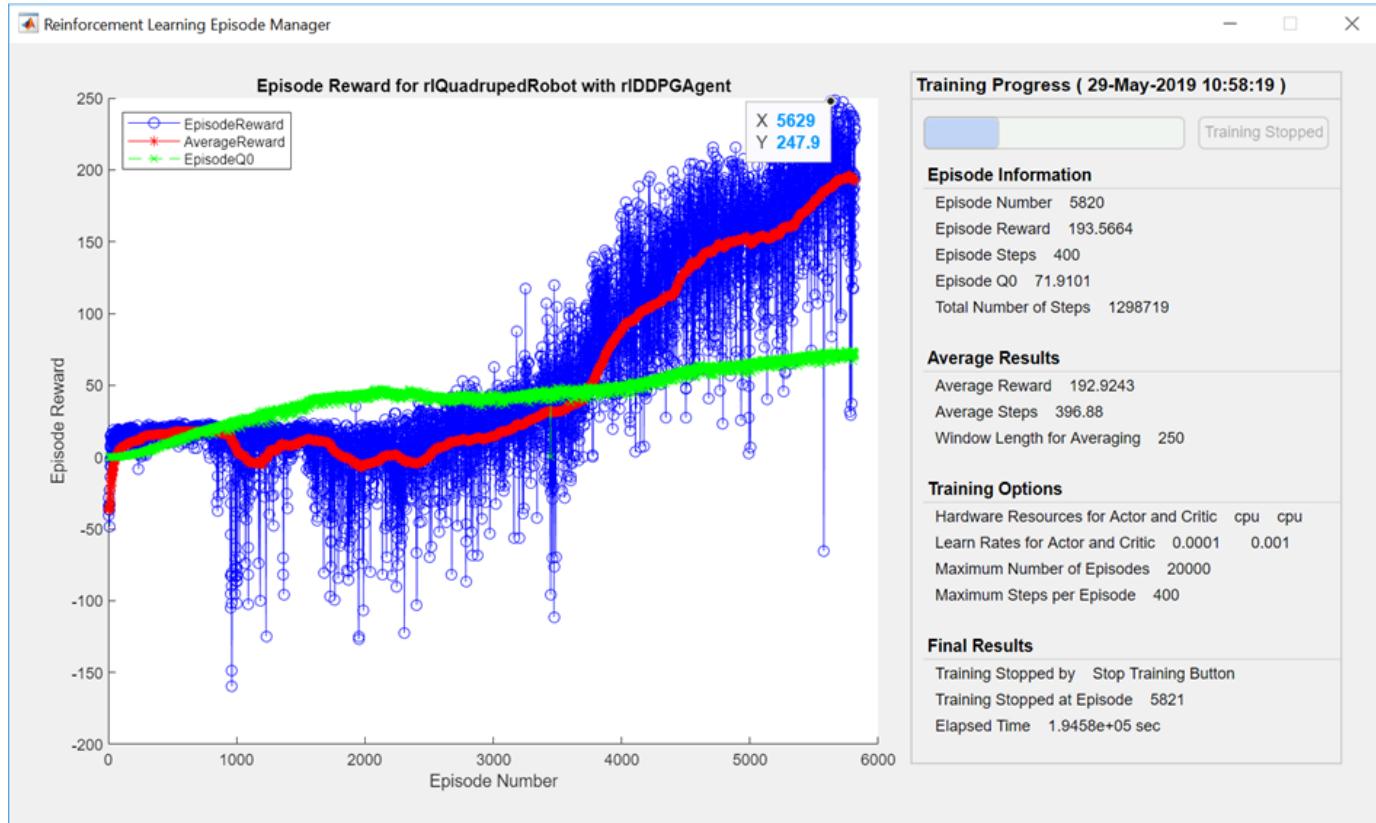
Train the agent using the `train` function. Due to the complexity of the robot model, this process is computationally intensive and takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`. Due to randomness of the parallel training, you can expect different training results from the plot below.

```

doTraining = false;
if doTraining
    % Train the agent
    trainingStats = train(agent,env,trainOpts);
else
    % Load pretrained agent for the example
    load('rlQuadrupedAgent.mat','agent')
end

```

Note that due to randomness in the parallel training, different training results are expected.



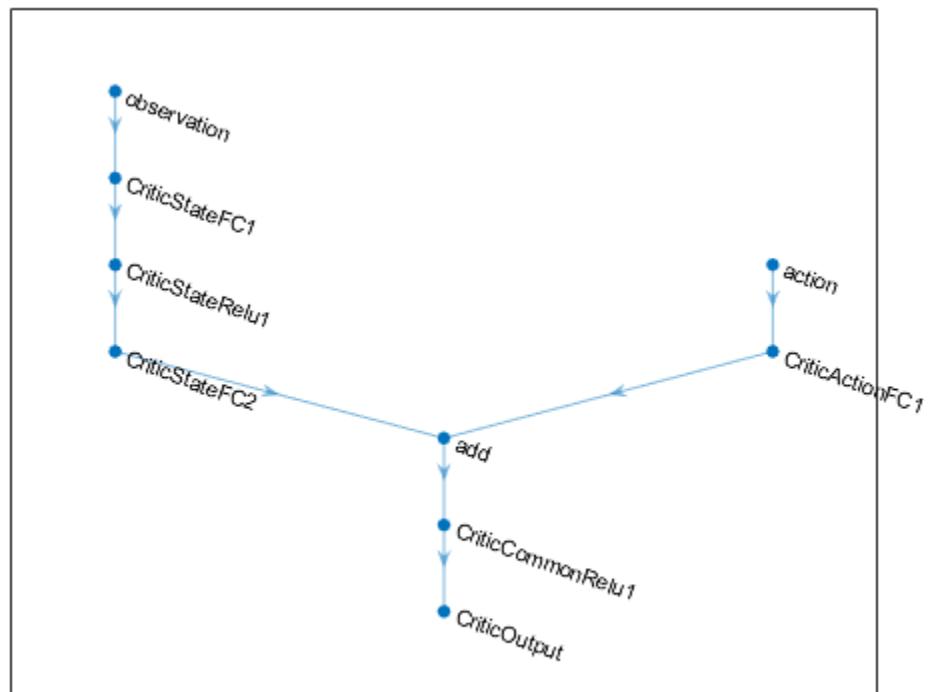
Simulate Trained Agent

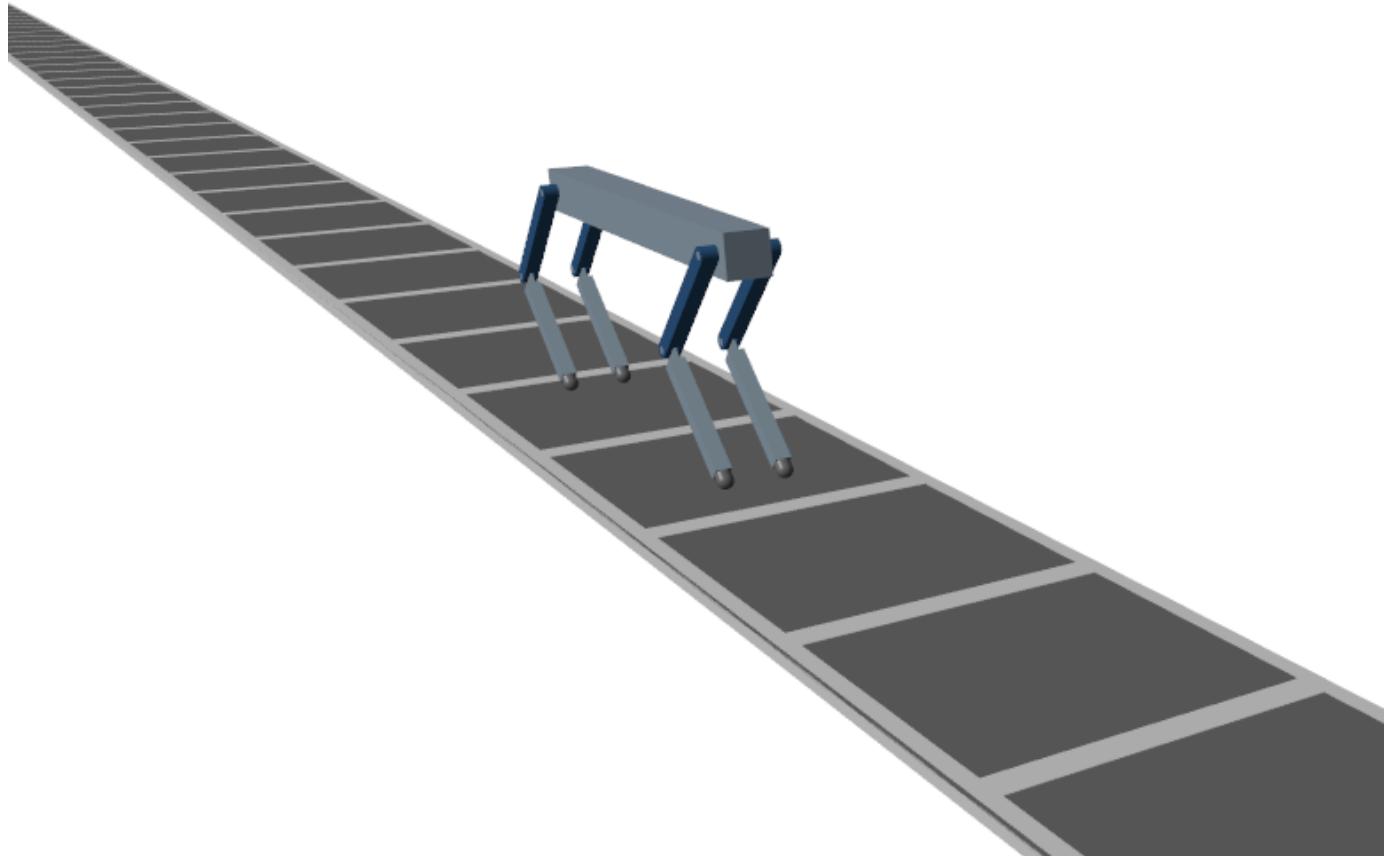
Fix the random generator seed for reproducibility.

```
rng(0)
```

To validate the performance of the trained agent, simulate it within the robot environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

```
simOptions = rlSimulationOptions('MaxSteps',maxSteps);
experience = sim(env,agent,simOptions);
```





References

[1] N. Heess et al, "Emergence of Locomotion Behaviours in Rich Environments," Technical Report, ArXiv, 2017.

[2] T.P. Lillicrap et al, "Continuous Control with Deep Reinforcement Learning," *International Conference on Learning Representations*, 2016.

See Also

[train](#)

More About

- "Reinforcement Learning Agents" on page 4-2
- "Train Reinforcement Learning Agents" on page 5-2
- "Define Reward Signals" on page 2-7

Train Custom LQR Agent

This example shows how to train a custom linear quadratic regulation (LQR) agent to control a discrete-time linear system modeled in MATLAB®.

Create Linear System Environment

The reinforcement learning environment for this example is a discrete-time linear system. The dynamics for the system are given by

$$x_{t+1} = Ax_t + Bu_t$$

The feedback control law is

$$u_t = -Kx_t$$

The control objective is to minimize the quadratic cost: $J = \sum_{t=0}^{\infty} (x_t'Qx_t + u_t'Ru_t)$.

In this example, the system matrices are

$$A = \begin{bmatrix} 1.05 & 0.05 & 0.05 \\ 0.05 & 1.05 & 0.05 \\ 0 & 0.05 & 1.05 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.1 & 0 & 0.2 \\ 0.1 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

$$A = [1.05, 0.05, 0.05; 0.05, 1.05, 0.05; 0, 0.05, 1.05];$$

$$B = [0.1, 0, 0.2; 0.1, 0.5, 0; 0, 0, 0.5];$$

The quadratic cost matrices are:

$$Q = \begin{bmatrix} 10 & 3 & 1 \\ 3 & 5 & 4 \\ 1 & 4 & 9 \end{bmatrix}$$

$$R = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

$$Q = [10, 3, 1; 3, 5, 4; 1, 4, 9];$$

$$R = 0.5 * eye(3);$$

For this environment, the reward at time t is given by $r_t = -x_t'Qx_t - u_t'Ru_t$, which is the negative of the quadratic cost. Therefore, maximizing the reward minimizes the cost. The initial conditions are set randomly by the reset function.

Create the MATLAB environment interface for this linear system and reward. The `myDiscreteEnv` function creates an environment by defining custom `step` and `reset` functions. For more information on creating such a custom environment, see “Create MATLAB Environment Using Custom Functions” on page 2-33.

$$\text{env} = \text{myDiscreteEnv}(A, B, Q, R);$$

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create Custom LQR Agent

For the LQR problem, the Q-function for a given control gain K can be defined as $Q_K(x, u) = \begin{bmatrix} x \\ u \end{bmatrix}^T H_K \begin{bmatrix} x \\ u \end{bmatrix}$,

where $H_K = \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix}$ is a symmetric, positive definite matrix.

The control law to maximize Q_K is $u = -(H_{uu})^{-1}H_{ux}x$, and the feedback gain is $K = (H_{uu})^{-1}H_{ux}$.

The matrix H_K contains $m = \frac{1}{2}n(n + 1)$ distinct element values, where n is the sum of the number of states and number of inputs. Denote θ as the vector corresponding to these m elements, where the off-diagonal elements in H_K are multiplied by two.

Represent the Q-function by θ , where θ contains the parameters to be learned.

$Q_K(x, u) = \theta'(K)\phi(x, u)$, where $\phi(x, u)$ is the quadratic basis function in terms of x and u .

The LQR agent starts with a stabilizing controller K_0 . To get an initial stabilizing controller, place the poles of the closed-loop system $A - BK_0$ inside the unit circle.

```
K0 = place(A,B,[0.4,0.8,0.5]);
```

To create a custom agent, you must create a subclass of the `rl.agent.CustomAgent` abstract class. For the custom LQR agent, the defined custom subclass is `LQRCustomAgent`. For more information, see “Custom Agents” on page 4-28. Create the custom LQR agent using Q , R , and K_0 . The agent does not require information on the system matrices A and B .

```
agent = LQRCustomAgent(Q,R,K0);
```

For this example, set the agent discount factor to one. To use a discounted future reward, set the discount factor to a value less than one.

```
agent.Gamma = 1;
```

Because the linear system has three states and three inputs, the total number of learnable parameters is $m = 21$. To ensure satisfactory performance of the agent, set the number of parameter estimates N_p to be greater than twice the number of learnable parameters. In this example, the value is $N_p = 45$.

```
agent.EstimateNum = 45;
```

To get good estimation results for θ , you must apply a persistently excited exploration model to the system. In this example, encourage model exploration by adding white noise to the controller output: $u_t = -Kx_t + e_t$. In general, the exploration model depends on the system models.

Train Agent

To train the agent, first specify the training options. For this example, use the following options.

- Run each training episode for at most 10 episodes, with each episode lasting at most 50 time steps.
- Display the command line display (set the **Verbose** option) and disable the training progress in the Episode Manager dialog box (set the **Plots** option).

For more information, see [rlTrainingOptions](#).

```
trainingOpts = rlTrainingOptions(...
    'MaxEpisodes',10, ...
    'MaxStepsPerEpisode',50, ...
    'Verbose',true, ...
    'Plots','none');
```

Train the agent using the **train** function.

```
trainingStats = train(agent,env,trainingOpts);
```

Episode:	1/ 10	Episode Reward :	-55.16	Episode Steps:	50	Avg Reward :	-55.16	Step Co
Episode:	2/ 10	Episode Reward :	-12.52	Episode Steps:	50	Avg Reward :	-33.84	Step Co
Episode:	3/ 10	Episode Reward :	-15.59	Episode Steps:	50	Avg Reward :	-27.76	Step Co
Episode:	4/ 10	Episode Reward :	-22.22	Episode Steps:	50	Avg Reward :	-26.37	Step Co
Episode:	5/ 10	Episode Reward :	-14.32	Episode Steps:	50	Avg Reward :	-23.96	Step Co
Episode:	6/ 10	Episode Reward :	-19.23	Episode Steps:	50	Avg Reward :	-16.78	Step Co
Episode:	7/ 10	Episode Reward :	-34.14	Episode Steps:	50	Avg Reward :	-21.10	Step Co
Episode:	8/ 10	Episode Reward :	-13.95	Episode Steps:	50	Avg Reward :	-20.77	Step Co
Episode:	9/ 10	Episode Reward :	-36.01	Episode Steps:	50	Avg Reward :	-23.53	Step Co
Episode:	10/ 10	Episode Reward :	-12.43	Episode Steps:	50	Avg Reward :	-23.15	Step Co

Simulate Agent and Compare with Optimal Solution

To validate the performance of the trained agent, simulate it within the MATLAB environment. For more information on agent simulation, see [rlSimulationOptions](#) and [sim](#).

```
simOptions = rlSimulationOptions('MaxSteps',20);
experience = sim(env,agent,simOptions);
totalReward = sum(experience.Reward)
```

totalReward = -20.1306

You can compute the optimal solution for the LQR problem using the **dlqr** function.

```
[Koptimal,P] = dlqr(A,B,Q,R);
```

The optimal reward is given by $J_{\text{optimal}} = -x_0'Px_0$.

```
x0 = experience.Observation.obs1.getdatasamples(1);
Joptimal = -x0'*P*x0;
```

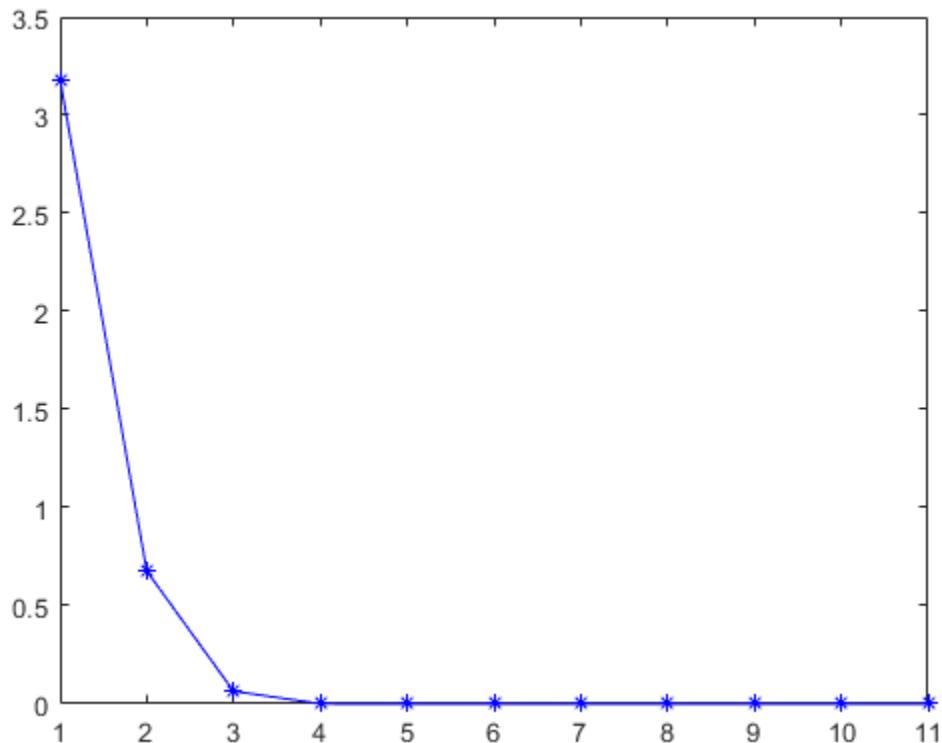
Compute the error in the reward between the trained LQR agent and the optimal LQR solution.

```
rewardError = totalReward - Joptimal
rewardError = 1.5270e-06
```

View the history of the 2-norm of error in the gains between the trained LQR agent and the optimal LQR solution.

```
% number of gain updates
len = agent.KUpdate;
```

```
err = zeros(len,1);
for i = 1:len
    % norm of error in the gain
    err(i) = norm(agent.KBuffer{i}-Koptimal);
end
plot(err, 'b*-')
```



Compute the norm of final error for the feedback gain.

```
gainError = norm(agent.K - Koptimal)
gainError = 2.2460e-11
```

Overall, the trained agent finds an LQR solution that is close to the true optimal LQR solution.

See Also

[train](#)

More About

- “Custom Agents” on page 4-28
- “Train Reinforcement Learning Agents” on page 5-2

Imitate MPC Controller for Lane Keep Assist

This example shows how to train, validate, and test a deep neural network that imitates the behavior of a model predictive controller for an automotive lane keeping assist system. It then compares the behavior of the deep neural network with that of the original controller.

Model predictive control (MPC) solves a constrained quadratic-programming (QP) optimization problem in real time based on the current state of the plant. Since MPC solves its optimization problem in an open-loop fashion, there is the potential to replace the controller with a trained deep neural network. Doing so is an appealing option, since evaluating a deep neural network can be more computationally efficient than solving a QP problem in real-time.

If the training of the network sufficiently traverses the state-space for the application, you can create a reasonable approximation of the controller behavior. You can then deploy the network for your control application. You can also use the network as a warm starting point for training the actor network of a reinforcement learning agent. For an example, see “Train DDPG Agent with Pretrained Actor Network” on page 5-165.

Design MPC Controller

Design an MPC controller for lane keeping assist. To do so, first create a dynamic model for the vehicle.

```
[sys,Vx] = createModelForMPCImLKA;
```

Create and design the MPC controller object `mpcobj`. Also, create an `mpcstate` object for setting the initial controller state. For details on the controller design, type `edit createMPCObjImLKA`.

```
[mpcobj,initialState] = createMPCObjImLKA(sys);
```

For more information on designing model predictive controllers for lane keeping assist applications, see “Lane Keeping Assist System Using Model Predictive Control” (Model Predictive Control Toolbox) and “Lane Keeping Assist with Lane Detection” (Model Predictive Control Toolbox).

Prepare Input Data

Load the input data from `InputDataFileImLKA.mat`. The columns of the data set contain:

- 1 Lateral velocity V_y
- 2 Yaw angle rate r
- 3 Lateral deviation e_1
- 4 Relative yaw angle e_2
- 5 Previous steering angle (control variable) u
- 6 Measured disturbance (road yaw rate: longitudinal velocity * curvature (ρ))
- 7 Cost function value
- 8 MPC iterations
- 9 Steering angle computed by MPC controller: u^*

The data in `InputDataFileImLKA.mat` was created by computing the MPC control action for randomly generated states, previous control actions, and measured disturbances. To generate your own training data, use the `collectDataImLKA` function.

Load the input data.

```
dataStruct = load('InputDataFileImLKA.mat');  
data = dataStruct.Data;
```

Divide the input data into training, validation, and testing data. First, determine number of validation data rows based on a given percentage.

```
totalRows = size(data,1);  
validationSplitPercent = 0.1;  
numValidationDataRows = floor(validationSplitPercent*totalRows);
```

Determine the number of test data rows based on a given percentage.

```
testSplitPercent = 0.05;  
numTestDataRows = floor(testSplitPercent*totalRows);
```

Randomly extract validation and testing data from the input data set. To do so, first randomly extract enough rows for both data sets.

```
randomIdx = randperm(totalRows,numValidationDataRows + numTestDataRows);  
randomData = data(randomIdx,:);
```

Divide the random data into validation and testing data.

```
validationData = randomData(1:numValidationDataRows,:);  
testData = randomData(numValidationDataRows + 1:end,:);
```

Extract the remaining rows as training data.

```
trainDataIdx = setdiff(1:totalRows,randomIdx);  
trainData = data(trainDataIdx,:);
```

Randomize the training data.

```
numTrainDataRows = size(trainData,1);  
shuffleIdx = randperm(numTrainDataRows);  
shuffledTrainData = trainData(shuffleIdx,:);
```

Reshape the training and validation data into 4-D matrices to be used with `trainNetwork`.

```
numObservations = 6;  
numActions = 1;  
  
trainInput = reshape(shuffledTrainData(:,1:6)',[numObservations 1 1 numTrainDataRows]);  
trainOutput = reshape(shuffledTrainData(:,9)',[numActions 1 1 numTrainDataRows]);  
  
validationInput = reshape(validationData(:,1:6)',[numObservations 1 1 numValidationDataRows]);  
validationOutput = reshape(validationData(:,9)',[numActions 1 1 numValidationDataRows]);  
validationCellArray = {validationInput,validationOutput};
```

Reshape the testing data to be used with `predict`.

```
testDataInput = reshape(testData(:,1:6)',[numObservations 1 1 numTestDataRows]);  
testDataOutput = testData(:,9');
```

Create Deep Neural Network

The deep neural network architecture uses the following types of layers.

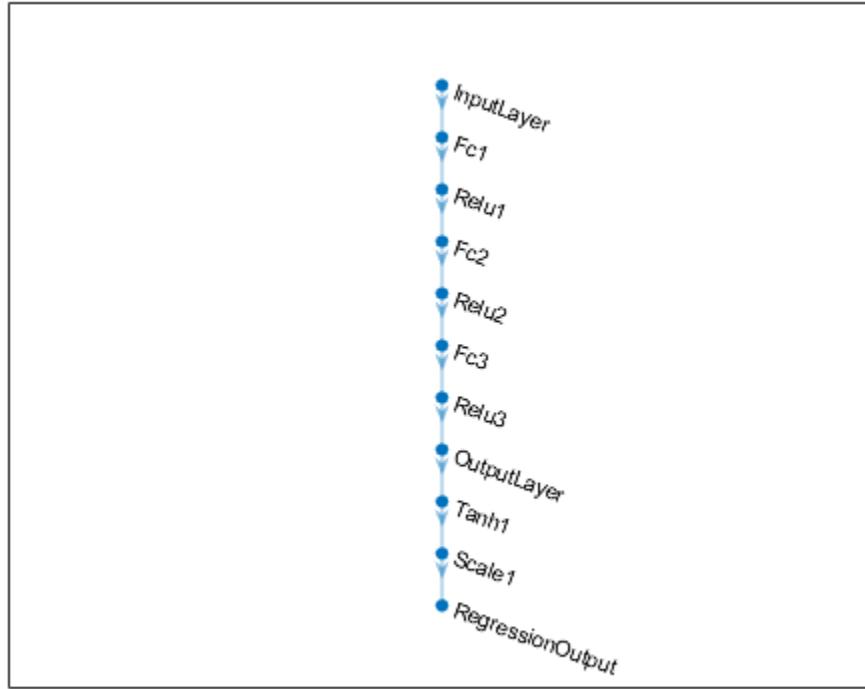
- `imageInputLayer` is input layer of the neural network.
- `fullyConnectedLayer` multiplies the input by a weight matrix and then adds a bias vector.
- `reluLayer` is the activation function of the neural network.
- `tanhLayer` constrains the value to the range to [-1,1].
- `scalingLayer` scales the value to the range to [-1.04,1.04], implies that the steering angle is constrained to be [-60,60] degrees.
- `regressionLayer` defines the loss function of the neural network.

Create the deep neural network that will imitate the MPC controller after training.

```
imitateMPCNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'InputLayer')
    fullyConnectedLayer(45, 'Name', 'Fc1')
    reluLayer('Name', 'Relu1')
    fullyConnectedLayer(45, 'Name', 'Fc2')
    reluLayer('Name', 'Relu2')
    fullyConnectedLayer(45, 'Name', 'Fc3')
    reluLayer('Name', 'Relu3')
    fullyConnectedLayer(numActions, 'Name', 'OutputLayer')
    tanhLayer('Name', 'Tanh1')
    scalingLayer('Name', 'Scale1', 'Scale', 1.04)
    regressionLayer('Name', 'RegressionOutput')
];
```

Plot the network.

```
plot(layerGraph(imitateMPCNetwork))
```



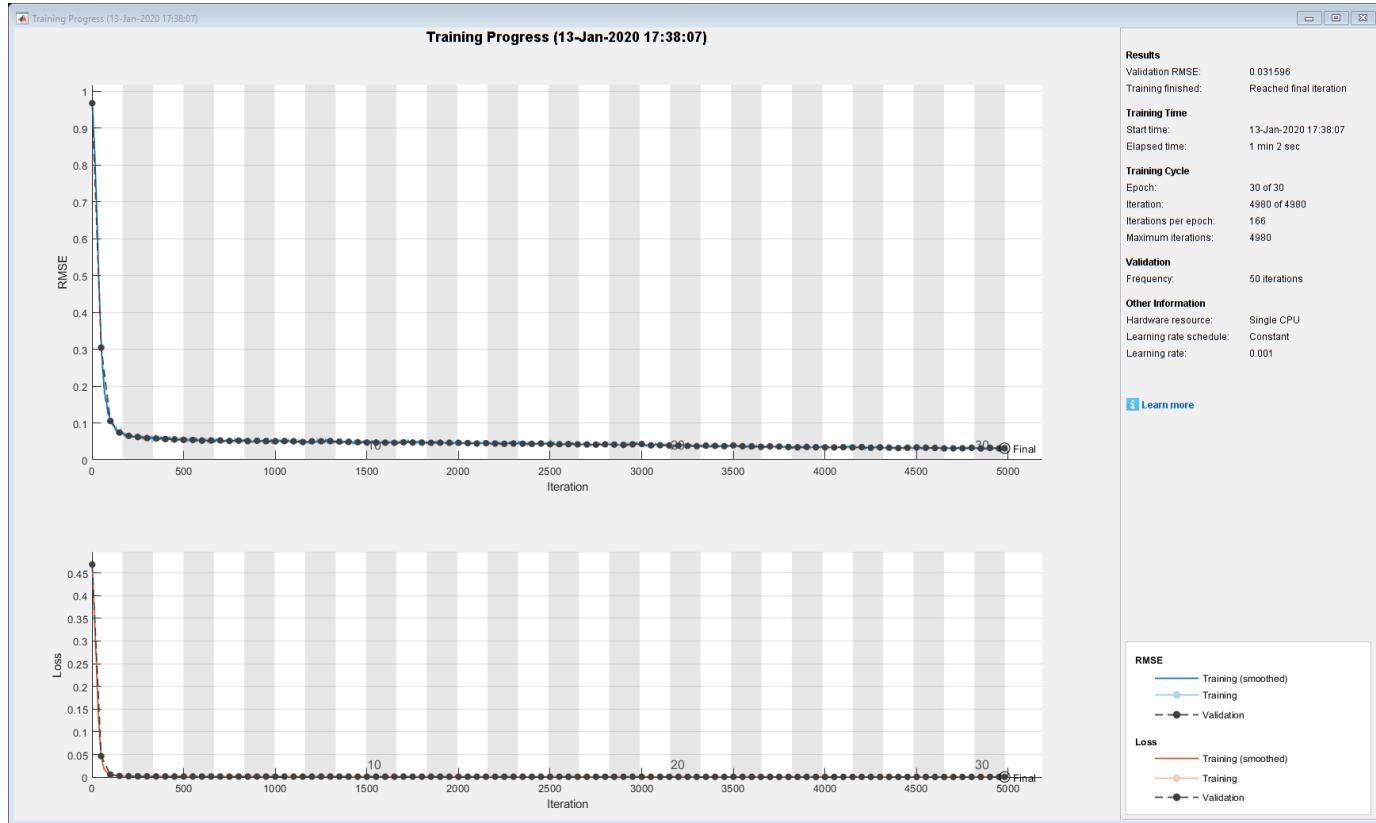
Train Deep Neural Network

Specify training options.

```
options = trainingOptions('adam', ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'Shuffle','every-epoch', ...
    'MaxEpochs', 30, ...
    'MiniBatchSize',512, ...
    'ValidationData',validationCellArray, ...
    'InitialLearnRate',1e-3, ...
    'GradientThresholdMethod','absolute-value', ...
    'ExecutionEnvironment','cpu', ...
    'GradientThreshold',10, ...
    'Epsilon',1e-8);
```

Train the deep neural network. To view detailed training information in the Command Window, you can set the '*Verbose*' training option to *true*.

```
imitateMPCNetObj = trainNetwork(trainInput,trainOutput,imitateMPCNetwork,options);
```



Training of the deep neural network stops when it reaches the final iteration.

The training and validation loss are nearly the same for each mini-batch indicating the trained network is not overfit.

Test Trained Network

Check that the trained deep neural network returns steering angles similar to the MPC controller control actions given the test input data. Compute the network output using the `predict` function.

```
predictedTestDataOutput = predict(imitateMPCNetObj,testDataInput);
```

Calculate the root mean-squared error between the network output and the testing data.

```
testRMSE = sqrt(mean(( testDataOutput - predictedTestDataOutput).^2));
fprintf('Test Data RMSE = %d\n', testRMSE);
```

Test Data RMSE = 3.195943e-02

The small RMSE value indicates that the network outputs closely reproduce the MPC controller outputs.

Compare Trained Network with MPC Controller

To compare the performance of the MPC controller and the trained deep neural network, run closed-loop simulations using the vehicle plant model.

Generate random initial conditions for the vehicle that are not part of the original input data set, with values selected from the following ranges:

- 1 lateral velocity V_y : range (-2,2) m/s
- 2 yaw angle rate r : range (-60,60) deg/s
- 3 lateral deviation e_1 : range (-1,1) m
- 4 relative yaw angle e_2 : range (-45,45) deg
- 5 last steering angle (control variable) u : range (-60,60) deg
- 6 measured disturbance (road yaw rate: longitudinal velocity * curvature (ρ)) : range (-0.01,0.01), minimum road radius: 100 m.

```
rng(5e7)
[x0,u0,rho] = generateRandomDataIMLKA(data);
```

Set the initial plant state and control action in the `mpcstate` object.

```
initialState.Plant = x0;
initialState.LastMove = u0;
```

Extract the sample time from the MPC controller. Also, set the number of simulation steps.

```
Ts = mpcoobj.Ts;
Tsteps = 30;
```

Obtain the A and B state-space matrices for the vehicle model.

```
A = sys.A;
B = sys.B;
```

Initialize the state and input trajectories for the MPC controller simulation.

```
xHistoryMPC = repmat(x0',Tsteps+1,1);
uHistoryMPC = repmat(u0',Tsteps,1);
```

Run a closed-loop simulation of the MPC controller and the plant using the `mpcmove` function.

```
for k = 1:Tsteps
    % Obtain plant output measurements, which correspond to the plant outputs.
    xk = xHistoryMPC(k,:)';
    % Compute the next control action using the MPC controller.
    uk = mpcoobj.mpcmove(mpcoobj,initialState,xk,zeros(1,4),Vx*rho);
    % Store the control action.
    uHistoryMPC(k,:) = uk;
    % Update the state using the control action.
    xHistoryMPC(k+1,:) = (A*xk + B*[uk;Vx*rho])';
end
```

Initialize the state and input trajectories for the deep neural network simulation.

```
xHistoryDNN = repmat(x0',Tsteps+1,1);
uHistoryDNN = repmat(u0',Tsteps,1);
lastMV = u0;
```

Run a closed-loop simulation of the trained network and the plant. The `neuralnetLKAmove` function computes the deep neural network output using the `predict` function.

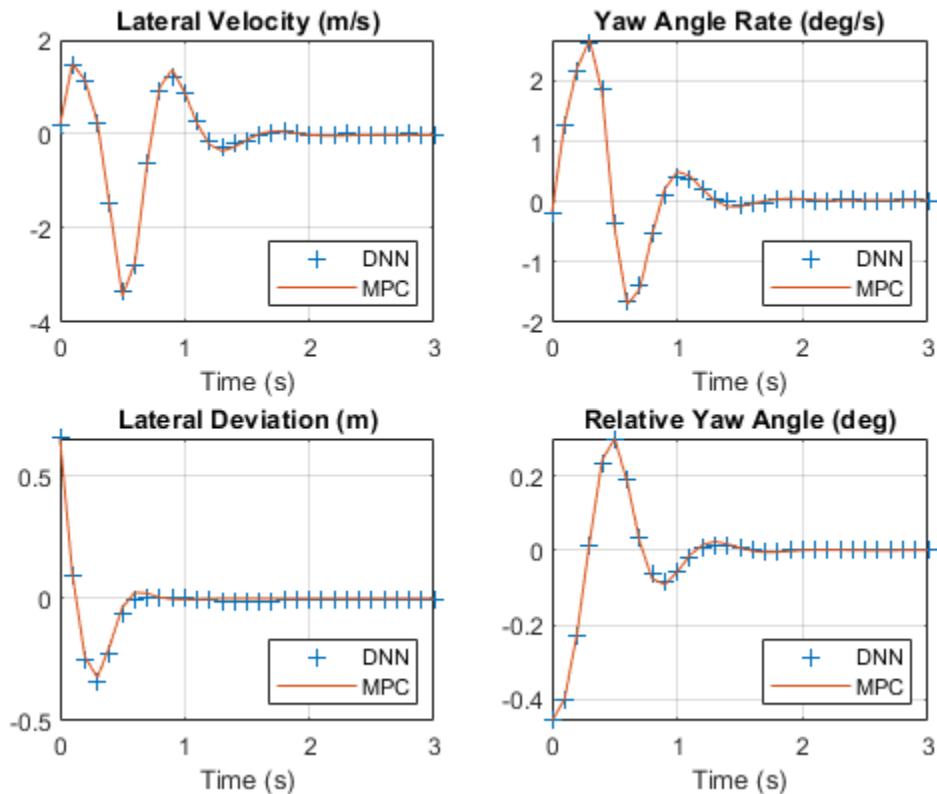
```

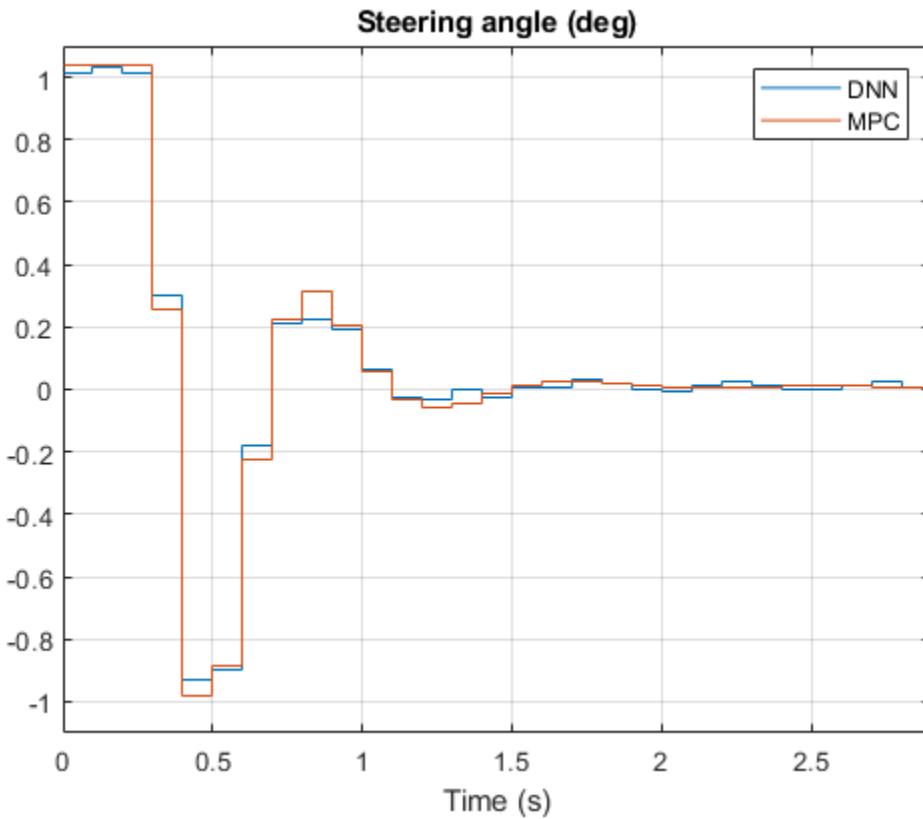
for k = 1:Tsteps
    % Obtain plant output measurements, which correspond to the plant outputs.
    xk = xHistoryDNN(k,:)';
    % Predict the next move using the trained deep neural network.
    uk = neuralnetLKAmove(imitateMPCNetObj,xk,lastMV,rho);
    % Store the control action and update the last MV for the next step.
    uHistoryDNN(k,:) = uk;
    lastMV = uk;
    % Update the state using the control action.
    xHistoryDNN(k+1,:) = (A*xk + B*[uk;Vx*rho])';
end

```

Plot the results, and compare the MPC controller and trained deep neural network (DNN) trajectories.

```
plotValidationResultsImLKA(Ts,xHistoryDNN,uHistoryDNN,xHistoryMPC,uHistoryMPC);
```





The neural network successfully imitates the behavior of the MPC controller. The vehicle state and control action trajectories for the controller and the deep neural network closely align.

See Also

`mpcmove` | `predict` | `trainNetwork`

More About

- “Lane Keeping Assist System Using Model Predictive Control” (Model Predictive Control Toolbox)
- “Lane Keeping Assist with Lane Detection” (Model Predictive Control Toolbox)

Train DDPG Agent with Pretrained Actor Network

This example shows how to train a deep deterministic policy gradient (DDPG) agent for lane keeping assist (LKA) in Simulink. To make training more efficient, the actor of the DDPG agent is initialized with a deep neural network that was previously trained using supervised learning. This actor trained is trained in the “Imitate MPC Controller for Lane Keep Assist” on page 5-157 example.

For more information on DDPG agents, see Deep Deterministic Policy Gradient Agents.

Simulink Model

The training goal for the lane-keeping application is to keep the ego vehicle traveling along the centerline of the a lane by adjusting the front steering angle. This example uses the same ego vehicle dynamics and sensor dynamics as the “Train DQN Agent for Lane Keeping Assist” on page 5-110 example.

```
m = 1575; % total vehicle mass (kg)
Iz = 2875; % yaw moment of inertia (mNs^2)
lf = 1.2; % longitudinal distance from center of gravity to front tires (m)
lr = 1.6; % longitudinal distance from center of gravity to rear tires (m)
Cf = 19000; % cornering stiffness of front tires (N/rad)
Cr = 33000; % cornering stiffness of rear tires (N/rad)
Vx = 15; % longitudinal velocity (m/s)
```

Define the sample time, T_s , and simulation duration, T , in seconds.

```
Ts = 0.1;
T = 15;
```

The output of the LKA system is the front steering angle of the ego vehicle. Considering the physical limitations of the ego vehicle, constrain its steering angle to the range [-60,60] degrees. Specify the constraints in radians.

```
u_min = -1.04;
u_max = 1.04;
```

Define the curvature of the road as a constant $0.001(m^{-1})$.

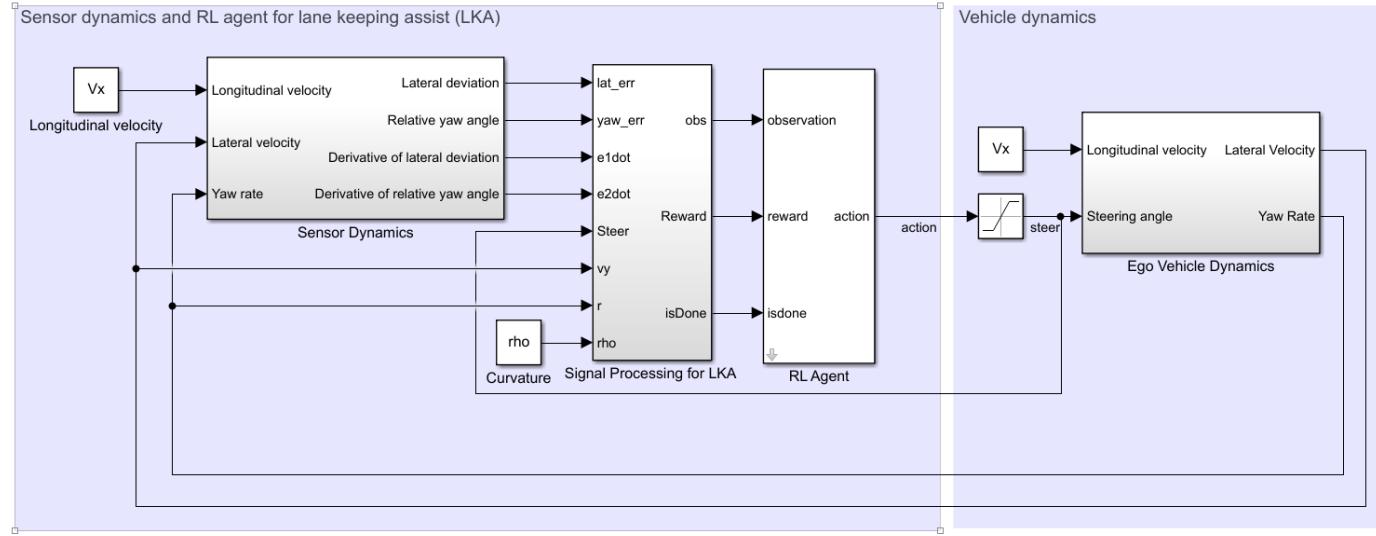
```
rho = 0.001;
```

Set initial values for the lateral deviation (`e1_initial`) and the relative yaw angle (`e2_initial`). During training, these initial conditions are set to random values for each training episode.

```
e1_initial = 0.2;
e2_initial = -0.1;
```

Open the model.

```
mdl = 'rlActorLKAMdl';
open_system(mdl)
```



Copyright 2018 The MathWorks, Inc.

Define the path to the RL Agent block within the model.

```
agentblk = [mdl '/RL Agent'];
```

Create Environment

Create a reinforcement learning environment interface for the ego vehicle. To do so, first define the observation and action specifications. These observations and actions are the same as the features for supervised learning used in “Imitate MPC Controller for Lane Keep Assist” on page 5-157.

The six observations for the environment are the lateral velocity v_y , yaw rate $\dot{\psi}$, lateral deviation e_1 , relative yaw angle e_2 , steering angle at previous step u_0 , and curvature ρ .

```
observationInfo = rlNumericSpec([6 1],...
    'LowerLimit', -inf*ones(6,1), 'UpperLimit', inf*ones(6,1));
observationInfo.Name = 'observations';
```

The action for the environment is the front steering angle. Specify the steering angle constraints when creating the action specification object.

```
actionInfo = rlNumericSpec([1 1], 'LowerLimit', u_min, 'UpperLimit', u_max);
actionInfo.Name = 'steering';
```

In the model, the Signal Processing for LKA block creates the observation vector signal, computes the reward function, and calculates the stop condition signal.

The reward r_t , provided at every time step t , is as follows, where u is the control input from the previous time step $t - 1$.

$$r_t = -(10e_1^2 + 5e_2^2 + 2u^2 + 5\dot{e}_1^2 + 5\dot{e}_2^2)$$

The simulation stops when $|e_1| > 1$.

Create the reinforcement learning environment.

```
env = rlSimulinkEnv(mdl,agentblk,observationInfo,actionInfo);
```

To define the initial condition for lateral deviation and relative yaw angle, specify an environment reset function using an anonymous function handle. The `localResetFcn` function, which is defined at the end of the example, sets the initial lateral deviation and relative yaw angle to random values.

```
env.ResetFcn = @(in)localResetFcn(in);
```

Fix the random generator seed for reproducibility.

```
rng(0)
```

Create DDPG Agent

A DDPG agent approximates the long-term reward given observations and actions using a critic value function representation. To create the critic, first create a deep neural network with two inputs, the state and action, and one output. For more information on creating a deep neural network value function representation, see “Create Policy and Value Function Representations” on page 3-2.

```
critic = createLaneKeepingCritic(observationInfo,actionInfo);
```

A DDPG agent decides which action to take given observations using an actor representation. To create the actor, first create a deep neural network with one input (the observation) and one output (the action).

```
[actor,actorOptions] = createLaneKeepingActor(observationInfo,actionInfo);
```

These initial actor and critic networks have random initial parameter values.

To create the DDPG agent, first specify the DDPG agent options.

```
agentOptions = rlDDPGAgentOptions(...
    'SampleTime',Ts, ...
    'TargetSmoothFactor',1e-3, ...
    'ExperienceBufferLength',1e6, ...
    'DiscountFactor',0.99, ...
    'MiniBatchSize',64 ...
);
agentOptions.NoiseOptions.Variance = 0.3;
agentOptions.NoiseOptions.VarianceDecayRate = 1e-5;
```

Create the DDPG agent using the specified actor representation, critic representation, and agent options. For more information, see `rlDDPGAgent`.

```
agent = rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

As a baseline, train the agent with an actor that has random initial parameters. To train the agent, first specify the training options. For this example, use the following options.

- Run each training episode for at most 50000 episodes, with each episode lasting at most 150 time steps.
- Display the training progress in the Episode Manager dialog box.
- Stop training when the episode reward reaches -1.
- Save a copy of the agent for each episode where the cumulative reward is greater than -2.5.

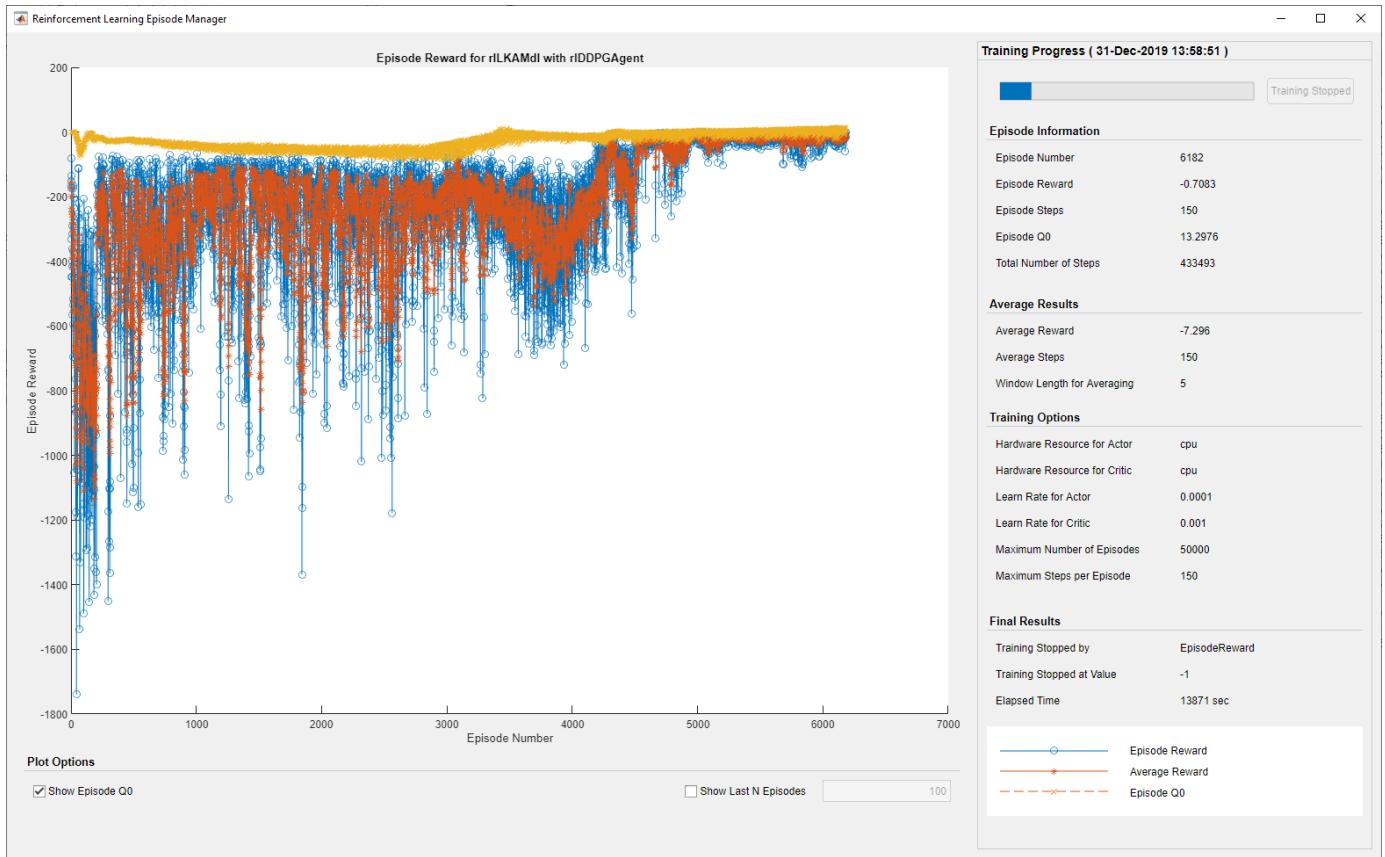
For more information, see `rlTrainingOptions`.

```
maxepisodes = 50000;
maxsteps = T/Ts;
trainingOpts = rlTrainingOptions(...
    'MaxEpisodes',maxepisodes, ...
    'MaxStepsPerEpisode',maxsteps, ...
    'Verbose',false, ...
    'Plots','training-progress', ...
    'StopTrainingCriteria','EpisodeReward', ...
    'StopTrainingValue',-1, ...
    'SaveAgentCriteria','EpisodeReward', ...
    'SaveAgentValue',-2.5);
```

Train the agent using the `train` function. Training is a computationally intensive process that takes several hours to complete. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
doTraining = false;

if doTraining
    % Train the agent.
    trainingStats = train(agent,env,trainingOpts);
else
    % Load pretrained agent for the example.
    load('savedParams_ddpg.mat');
    agent = loadPretrainedParams(agent,actorParams,criticParams);
end
```



Train Agent with Pretrained Actor

You can set the actor network of your agent to a deep neural network that has been previously trained. For this example, use the deep neural network from the “Imitate MPC Controller for Lane Keep Assist” on page 5-157 example. This network was trained to imitate a model predictive controller using supervised learning.

Load the pretrained actor network.

```
load('imitateMPCNetActorObj.mat','imitateMPCNetObj');
```

Create an actor representation using the pretrained actor.

```
supervisedActor = rlDeterministicActorRepresentation(imitateMPCNetObj,observationInfo,actionInfo
'Observation',imitateMPCNetObj.InputNames,'Action',{'Scale1'},actorOptions);
```

Check that loaded the deep neural network is the same as the actor for the DDPG agent. To do so, evaluate both the network and the agent using the same random input observation.

```
 testData = rand(6,1);
```

Evaluate the deep neural network.

```
 predictIMNN = predict(imitateMPCNetObj,testData);
```

Evaluate the actor.

```
 evaluateRLRep = getAction(supervisedActor,{testData});
```

5 Train and Validate Agents

Compare the results.

```
error = evaluateRLRep{} - predictImNN  
error = single  
    0
```

Create a DDPG agent using the pretrained actor.

```
agent = rlDDPGAgent(supervisedActor,critic,agentOptions);
```

Reduce the maximum number of training episodes and train the agent using the `train` function. To save time while running this example, load a pretrained agent by setting `doTraining` to `false`. To train the agent yourself, set `doTraining` to `true`.

```
trainingOpts.MaxEpisodes = 5000;  
doTraining = false;  
  
if doTraining  
    % Train the agent.  
    trainingStats = train(agent,env,trainingOpts);  
else  
    % Load pretrained agent for the example.  
    load('savedParams_imActor.mat');  
    agent = loadPretrainedParams(agent,actorParams,criticParams);  
end
```



By using the pretrained actor network, the training of the DDPG agent is more efficient. Both the total training time and the total number of training steps have improved by approximately 20%. Also, the number of episodes for the training to approach the neighborhood of the optimal result decreased from approximately 4500 to approximately 3500.

Simulate DDPG Agent

To validate the performance of the trained agent, uncomment the following two lines and simulate it within the environment. For more information on agent simulation, see `rlSimulationOptions` and `sim`.

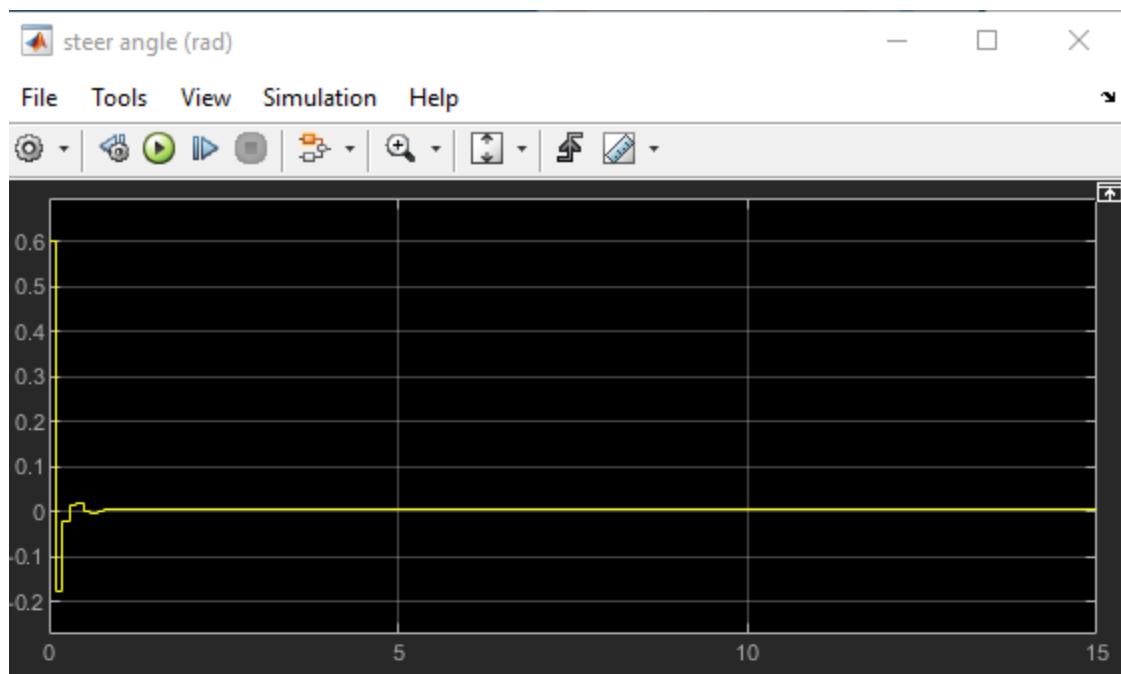
```
% simOptions = rlSimulationOptions('MaxSteps',maxsteps);  
% experience = sim(env,agent,simOptions);
```

To check the performance of the trained agent within the Simulink model, simulate the model using the previously defined initial conditions (`e1_initial = 0.2` and `e2_initial = -0.1`).

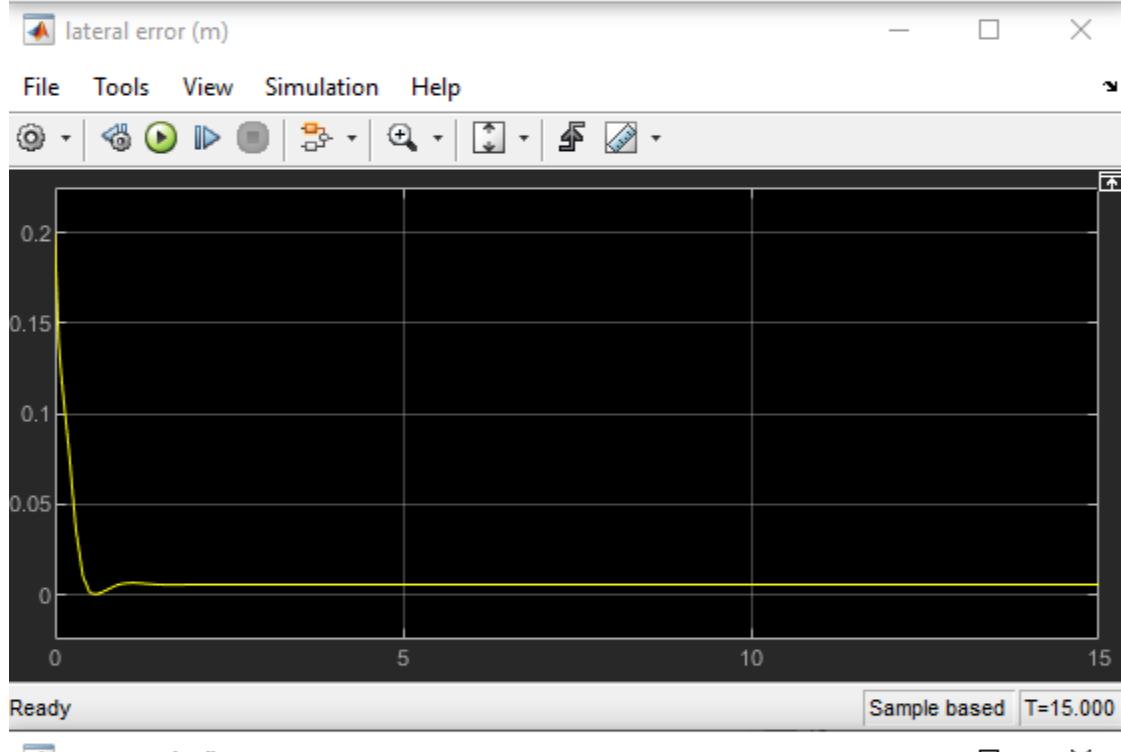
```
sim(mdl)
```

As shown below, the lateral error (middle plot) and relative yaw angle (bottom plot) are both driven to zero. The vehicle starts with a lateral deviation from the centerline (0.2 m) and a nonzero yaw angle error (-0.1 rad). The lane-keeping controller makes the ego vehicle travel along the centerline after around two seconds. The steering angle (top plot) shows that the controller reaches steady state after about two seconds.

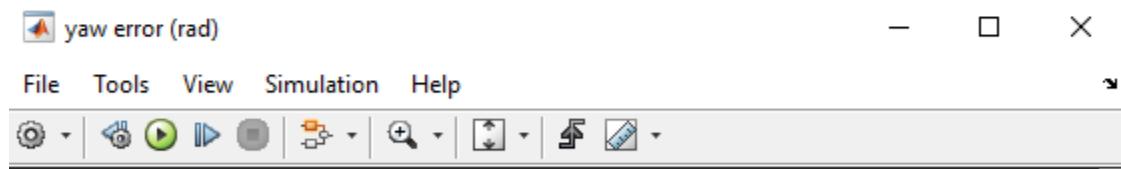
5 Train and Validate Agents



Ready Sample based T=15.000



Ready Sample based T=15.000



Close the Simulink model.

```
bdclose(mdl)
```

Local Functions

```
function in = localResetFcn(in)
% Set random value for lateral deviation.
in = setVariable(in,'e1_initial', 0.5*(-1+2*rand));

% Set random value for relative yaw angle.
in = setVariable(in,'e2_initial', 0.1*(-1+2*rand));
end

function agent = loadPretrainedParams(agent,actorParams,criticParams)
% Set actor parameters.
actor = getActor(agent);
pretrainedActor = setLearnableParameters(actor,actorParams);

% Set critic parameters.
critic = getCritic(agent);
pretrainedCritic = setLearnableParameters(critic,criticParams);

% Set actor and critic representations in the agent.
agent = setActor(agent,pretrainedActor);
agent = setCritic(agent,pretrainedCritic);
end
```

See Also

[rlDDPGAgent | train](#)

More About

- “Imitate MPC Controller for Lane Keep Assist” on page 5-157
- “Train DQN Agent for Lane Keeping Assist” on page 5-110
- “Deep Deterministic Policy Gradient Agents” on page 4-14

Imitate Nonlinear MPC Controller for Flying Robot

This example shows how to train, validate, and test a deep neural network (DNN) that imitates the behavior of a nonlinear model predictive controller for a flying robot. It then compares the behavior of the deep neural network with that of the original controller. To train the deep neural network, this example uses the data aggregation (DAgger) approach as in [1].

Nonlinear model predictive control (NLMPc) solves a constrained nonlinear optimization problem in real time based on the current state of the plant. Since NLMPc solves its optimization problem in an open-loop fashion, there is the potential to replace the controller with a trained DNN. Doing so is an appealing option, since evaluating a DNN can be more computationally efficient than solving a nonlinear optimization problem in real-time.

If the training of the DNN creates a reasonable approximation of the controller behavior, you can then deploy the network for your control application. You can also use the network as a warm starting point for training the actor network of a reinforcement learning agent. For an example that does so with a DNN trained for an MPC application, see “Train DDPG Agent with Pretrained Actor Network” on page 5-165.

Design Nonlinear MPC Controller

Design a nonlinear MPC controller for a flying robot. The dynamics for the flying robot are the same as in “Trajectory Optimization and Control of Flying Robot Using Nonlinear MPC” (Model Predictive Control Toolbox) example. First, define the limit for the control variables, which are the robot thrust levels.

```
umax = 3;
```

Create the nonlinear MPC controller object `nlobj`. To reduce command-window output, disable the MPC update messages.

```
mpcverbosity off;
nlobj = createMPCobjImFlyingRobot(umax);
```

Prepare Input Data

Load the input data from `DAggerInputDataFileImFlyingRobot.mat`. The columns of the data set contain:

- 1** x is the position of the robot along the x-axis.
- 2** y is the position of the robot along the y-axis.
- 3** θ is the orientation of the robot.
- 4** \dot{x} is the velocity of the robot along the x-axis.
- 5** \dot{y} is the velocity of the robot along the y-axis.
- 6** $\dot{\theta}$ is the angular velocity of the robot.
- 7** u_l is the thrust on the left side of the flying robot
- 8** u_r is the thrust on the right side of the flying robot
- 9** u_l^* is the thrust on the left side computed by NLMPc
- 10** u_r^* is the thrust on the right side computed by NLMPc

The data in `DAggerInputDataFileImFlyingRobot.mat` is created by computing the NLMPC control action for randomly generated states ($x, y, \theta, \dot{x}, \dot{y}, \dot{\theta}$), and previous control actions (u_l, u_r). To generate your own training data, use the `collectDataImFlyingRobot` function.

Load the input data.

```
fileName = 'DaggerInputDataFileImFlyingRobot.mat';
DaggerData = load(fileName);
data = DaggerData.data;
existingData = data;
numCol = size(data,2);
```

Create Deep Neural Network

The deep neural network architecture uses the following types of layers.

- `imageInputLayer` is the input layer of the neural network.
- `fullyConnectedLayer` multiplies the input by a weight matrix and then adds a bias vector.
- `reluLayer` is the activation function of the neural network.
- `tanhLayer` constrains the value to the range to [-1,1].
- `scalingLayer` scales the value to the range to [-3,3].
- `regressionLayer` defines the loss function of the neural network.

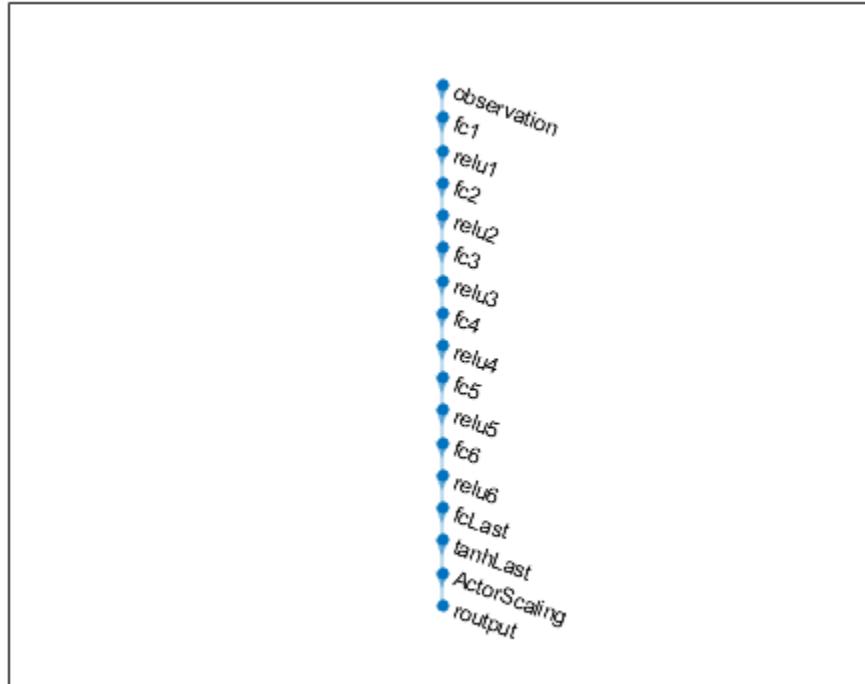
Create the deep neural network that will imitate the NLMPC controller after training.

```
numObservations = numCol-2;
numActions = 2;
hiddenLayerSize = 256;

imitateMPCNetwork = [
    imageInputLayer([numObservations 1 1], 'Normalization', 'none', 'Name', 'observation')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc2')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc3')
    reluLayer('Name', 'relu3')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc4')
    reluLayer('Name', 'relu4')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc5')
    reluLayer('Name', 'relu5')
    fullyConnectedLayer(hiddenLayerSize, 'Name', 'fc6')
    reluLayer('Name', 'relu6')
    fullyConnectedLayer(numActions, 'Name', 'fcLast')
    tanhLayer('Name', 'tanhLast')
    scalingLayer('Name', 'ActorScaling', 'Scale', umax)
    regressionLayer('Name', 'routput')];
```

Plot the network.

```
plot(layerGraph(imitateMPCNetwork))
```



Behavior Cloning Approach

One approach to learning an expert policy using supervised learning is the behavior cloning method. This method divides the expert demonstrations (NLMPC control actions in response to observations) into state-action pairs and applies supervised learning to train the network.

Specify training options.

```
% initialize validation cell array
validationCellArray = {0,0};

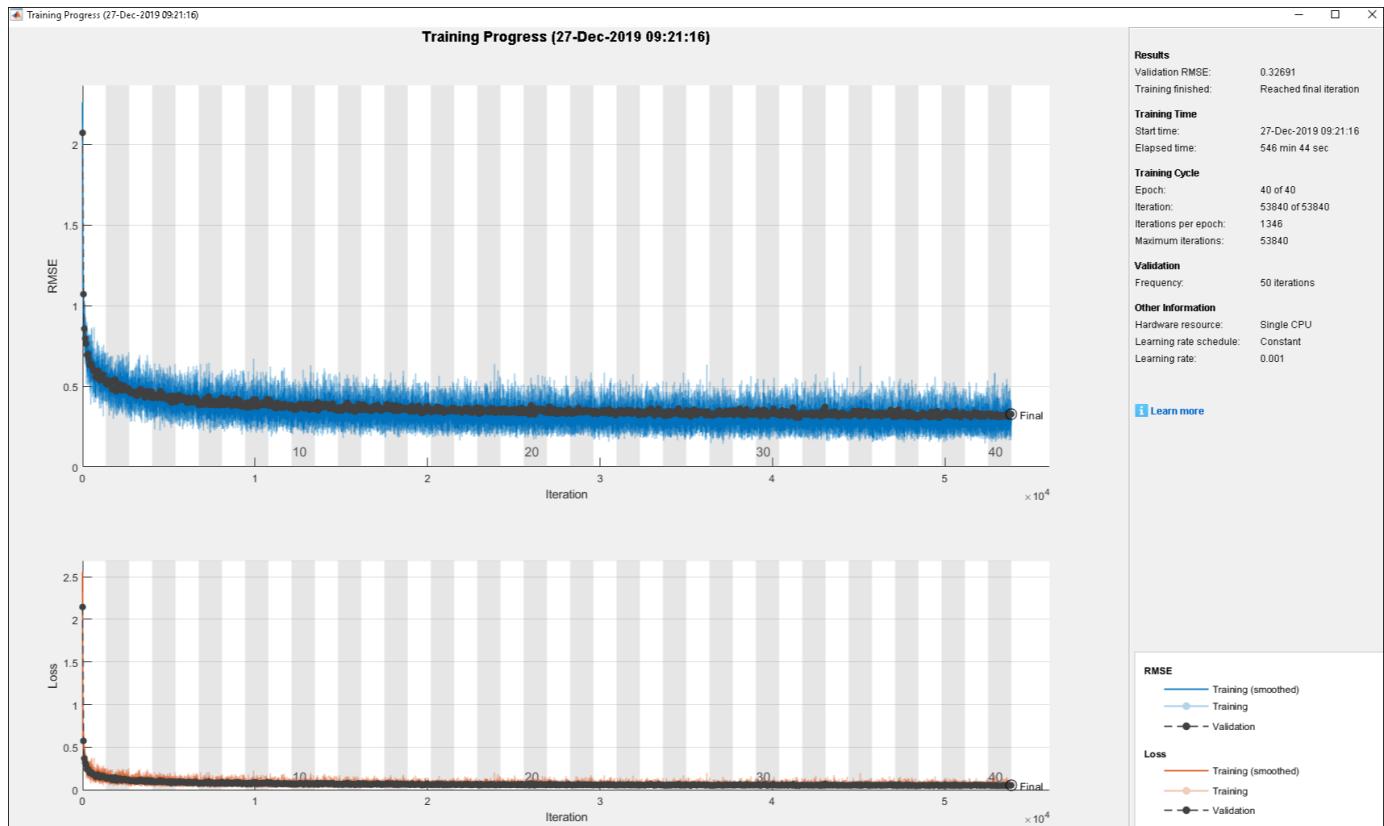
options = trainingOptions('adam', ...
    'Verbose', false, ...
    'Plots', 'training-progress', ...
    'Shuffle', 'every-epoch', ...
    'MiniBatchSize', 512, ...
    'ValidationData', validationCellArray, ...
    'InitialLearnRate', 1e-3, ...
    'ExecutionEnvironment', 'cpu', ...
    'GradientThreshold', 10, ...
    'MaxEpochs', 40 ...
);
```

You can train the behavior cloning neural network by following below steps

- 1 Collect data using the `collectDataImFlyingRobot` function.
- 2 Train the behavior cloning network using the `behaviorCloningTrainNetwork` function.

Training a DNN is a computationally intensive process. To save time, load a pretrained neural network object.

```
load('behaviorCloningMPCImDNNObject.mat');
imitateMPCNetBehaviorCloningObj = behaviorCloningNNObj.imitateMPCNetObj;
```



The training of the DNN using behavior cloning reduces the gap between the DNN and NLMPC performance. However, the behavior cloning neural network fails to imitate the behavior of the NLMPC controller correctly on some randomly generated data.

Data Aggregation Approach

To improve the performance of the DNN, you can learn the policy using an interactive demonstrator method. DAgger is an iterative method where the DNN is run in the closed-loop environment. The expert, in this case the NLMPC controller, outputs actions based on the states visited by the DNN. In this manner, more training data is aggregated and the DNN is retrained for improved performance. For more information, see [1].

Train the deep neural network using the `DAggerTrainNetwork` function. It creates `DAggerImFlyingRobotDNNObj.mat` file that contains the following information.

- `DatasetPath`: path where the dataset corresponding to each iteration is stored
- `policyObjs`: policies that were trained in each iteration
- `finalData`: total training data collected till final iteration
- `finalPolicy`: best policy among all the collected policies

First, create and initialize the parameters for training. Use the network trained using behavior cloning (`imitateMPCNetBehaviorCloningObj`) as the starting point for the DAgger training.

```
[dataStruct,nlmpcStruct,tuningParamsStruct,neuralNetStruct] = loadDAggerParameters(existingData,  
    numCol,nlobj,uMax,options,imitateMPCNetBehaviorCloningObj);
```

To save time, load a pretrained neural network by setting `doTraining` to `false`. To train the DAgger yourself, set `doTraining` to `true`.

```
doTraining = false;  
  
if doTraining  
    DAgger = DAggerTrainNetwork(nlmpcStruct,dataStruct,neuralNetStruct,tuningParamsStruct);  
else  
    load('DAggerImFlyingRobotDNNObj.mat');  
end  
DNN = DAgger.finalPolicy;
```

As an alternative, you can train the neural network with a modified policy update rule using the `DAggerModifiedTrainNetwork` function. In this function, after every 20 training iterations, the DNN is set to the most optimal configuration from the previous 20 iterations. To run this example with a neural network object with the modified DAgger approach, use the `DAggerModifiedImFlyingRobotDNNObj.mat` file.

Compare Trained DAgger Network with NLMP Controller

To compare the performance of the NLMP controller and the trained DNN, run closed-loop simulations with the flying robot model.

Set initial condition for the states of the flying robot ($x, y, \theta, \dot{x}, \dot{y}, \dot{\theta}$) and the control variables of flying robot (u_l, u_r).

```
x0 = [-1.8200 0.5300 -2.3500 1.1700 -1.0400 0.3100]';  
u0 = [-2.1800 -2.6200]';
```

Run a closed-loop simulation of the NLMP controller.

```
% Duration  
Tf = 15;  
% Sample time  
Ts = nlobj.Ts;  
% Simulation steps  
Tsteps = Tf/Ts+1;  
% Run NLMP in closed loop.  
tic  
[xHistoryMPC,uHistoryMPC] = simModelMPCImFlyingRobot(x0,u0,nlobj,Tf);  
toc
```

Elapsed time is 86.518379 seconds.

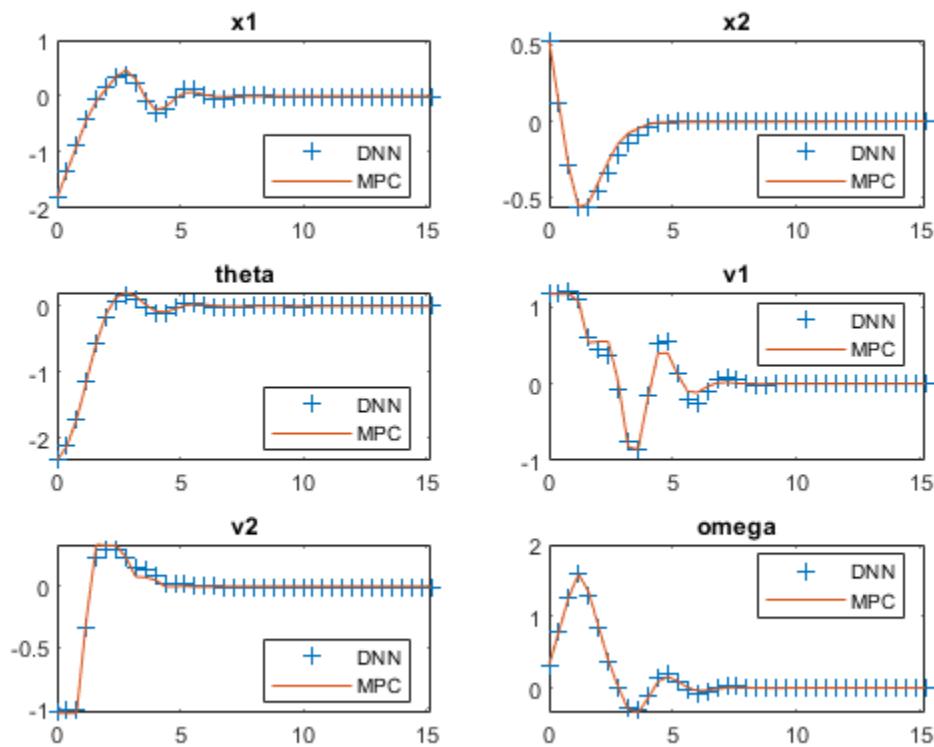
Run a closed-loop simulation of the trained DAgger network.

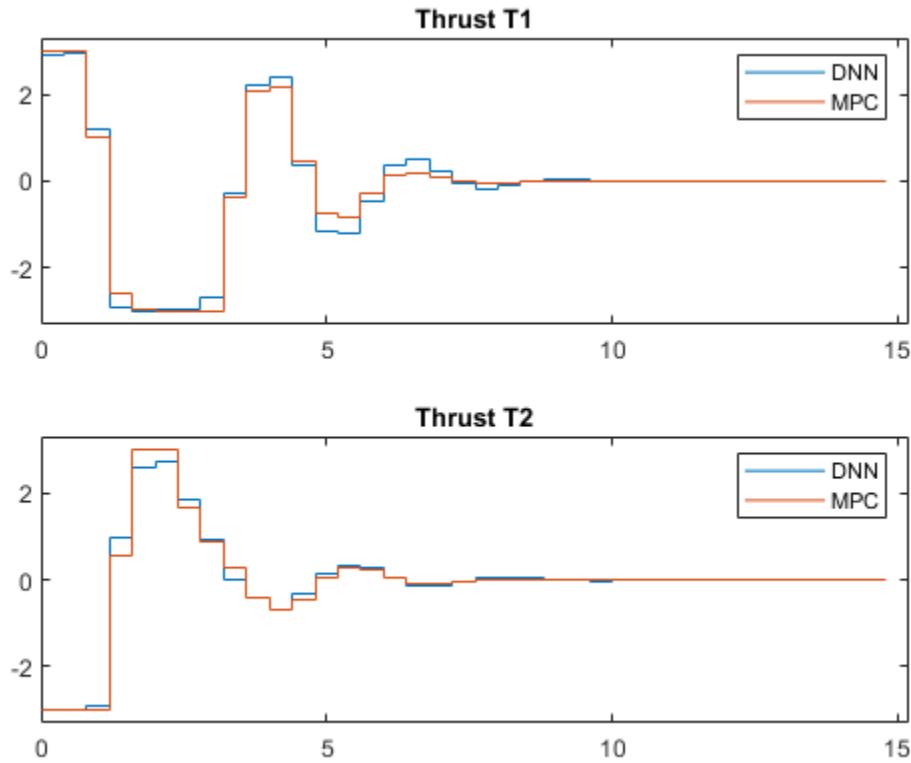
```
tic  
[xHistoryDNN,uHistoryDNN] = simModelDAggerImFlyingRobot(x0,u0,DNN,Ts,Tf);  
toc
```

Elapsed time is 2.727963 seconds.

Plot the results, and compare the NLMPC and trained DNN trajectories.

```
plotSimResultsImFlyingRobot(nlobj,xHistoryMPC,uHistoryMPC,xHistoryDNN,uHistoryDNN,umax,Tf)
```



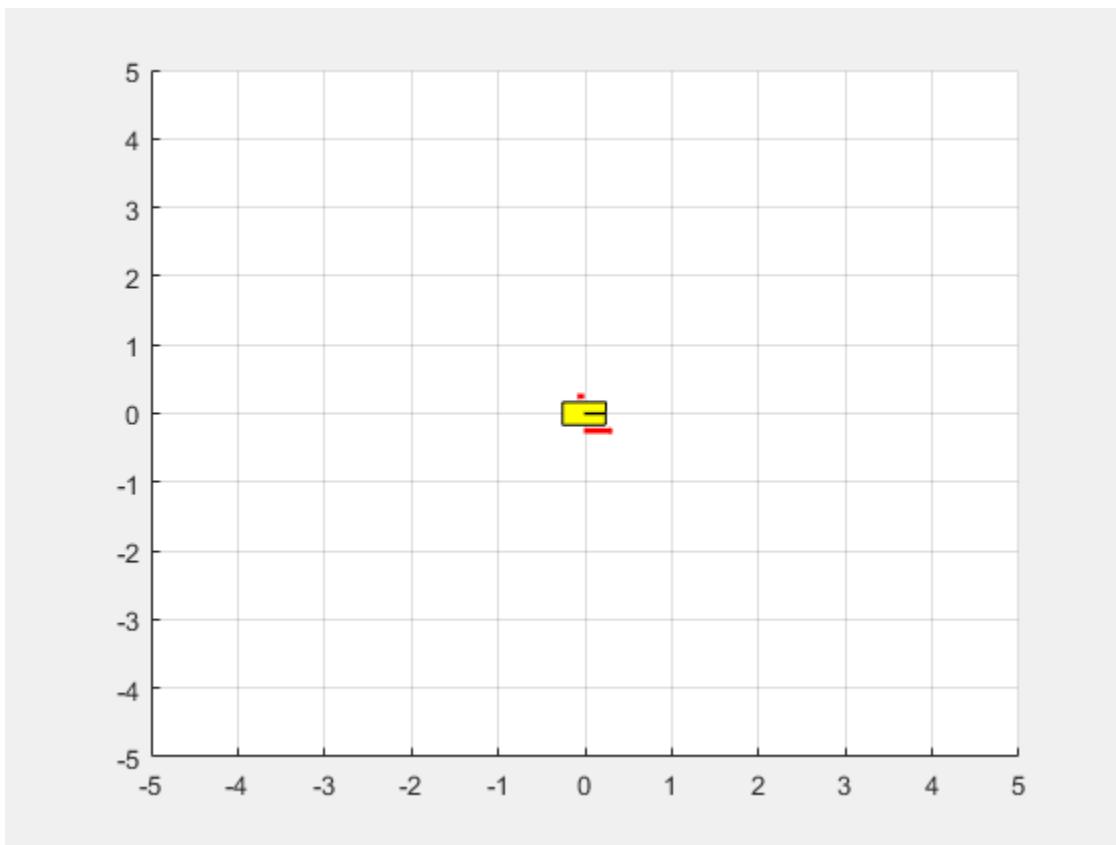


The DAgger neural network successfully imitates the behavior of the NLMPc controller. The flying robot states and control action trajectories for the controller and the DAgger deep neural network closely align. The closed-loop simulation time for the DNN is significantly less than that of the NLMPc controller.

Animate the Flying Robot with Trained DAgger Network

To validate the performance of the trained DNN, animate the flying robot with data from the DNN closed-loop simulation. The flying robot lands at the origin successfully.

```
Lx = 5;
Ly = 5;
for ct = 1:Tsteps
    x = xHistoryDNN(ct,1);
    y = xHistoryDNN(ct,2);
    theta = xHistoryDNN(ct,3);
    tL = uHistoryDNN(ct,1);
    tR = uHistoryDNN(ct,2);
    rl.env.viz.plotFlyingRobot(x,y,theta,tL,tR,Lx,Ly);
    pause(0.05);
end
```



```
% Turn on MPC messages  
mpcverbosity on;
```

References

- [1] Osa, Takayuki, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. 'An Algorithmic Perspective on Imitation Learning'. *Foundations and Trends in Robotics* 7, no. 1–2 (2018): 1–179. <https://doi.org/10.1561/2300000053>.

See Also

`rlDDPGAgent | train`

More About

- "Imitate MPC Controller for Lane Keep Assist" on page 5-157
- "Train DQN Agent for Lane Keeping Assist" on page 5-110
- "Deep Deterministic Policy Gradient Agents" on page 4-14

Train Reinforcement Learning Policy Using Custom Training Loop

This example shows how to define a custom training loop for a reinforcement learning policy. You can use this workflow to train reinforcement learning policies with your own custom training algorithms rather than using one of the built-in agents from the Reinforcement Learning Toolbox™ software.

Using this workflow, you can train policies that use any of the following policy and value function representations.

- `rlStochasticActorRepresentation` — Stochastic actor representation
- `rlDeterministicActorRepresentation` — Deterministic actor representation
- `rlValueRepresentation` — Value function critic representation
- `rlQValueRepresentation` — Q-Value function critic representation

In this example, a stochastic actor policy with a discrete action space is trained using the REINFORCE algorithm (with no baseline). For more information on the REINFORCE algorithm, see “Policy Gradient Agents” on page 4-11.

Fix the random generator seed for reproducibility.

```
rng(0)
```

For more information on the functions you can use for custom training, see Functions for Custom Training on page 5-0 .

Environment

For this example, a reinforcement learning policy is trained in a discrete cart-pole environment. The objective in this environment is to balance the pole by applying forces (actions) on the cart. Create the environment using the `rlPredefinedEnv` function.

```
env = rlPredefinedEnv('CartPole-Discrete');
```

Extract the observation and action specifications from the environment.

```
obsInfo = getObservationInfo(env);  
actInfo = getActionInfo(env);
```

Obtain the number of observations (`numObs`) and actions (`numAct`).

```
numObs = obsInfo.Dimension(1);  
numAct = actInfo.Dimension(1);
```

For more information on this environment, see “Load Predefined Control System Environments” on page 2-15.

Policy

The reinforcement learning policy in this example is a discrete-action stochastic policy. It is represented by a deep neural network that contains `fullyConnectedLayer`, `reluLayer`, and `softmaxLayer` layers. This network outputs probabilities for each discrete action given the current observations. The `softmaxLayer` ensures that the representation outputs probability values in the range [0 1] and that all probabilities sum to 1.

Create the deep neural network for the actor.

```
actorNetwork = [imageInputLayer([numObs 1 1], 'Normalization', 'none', 'Name', 'state')
    fullyConnectedLayer(24, 'Name', 'fc1')
    reluLayer('Name', 'relu1')
    fullyConnectedLayer(24, 'Name', 'fc2')
    reluLayer('Name', 'relu2')
    fullyConnectedLayer(2, 'Name', 'output')
    softmaxLayer('Name', 'actionProb')];
```

Create the actor representation using an `rlStochasticActorRepresentation` object.

```
actorOpts = rlRepresentationOptions('LearnRate', 1e-3, 'GradientThreshold', 1);
actor = rlStochasticActorRepresentation(actorNetwork, ...
    obsInfo, actInfo, 'Observation', 'state', actorOpts);
```

For this example, the loss function for the policy is implemented in `actorLossFunction` on page 5-0 .

Set the loss function using the `setLoss` function.

```
actor = setLoss(actor, @actorLossFunction);
```

Training Setup

Configure the training to use the following options:

- Set up the training to last at most 5000 episodes, with each episode lasting at most 250 steps.
- To calculate the discounted reward, choose a discount factor of 0.995.
- Terminate the training after the maximum number of episodes is reached or when the average reward across 100 episodes reaches the value of 220.

```
numEpisodes = 5000;
maxStepsPerEpisode = 250;
discountFactor = 0.995;
aveWindowSize = 100;
trainingTerminationValue = 220;
```

Create a vector for storing the cumulative reward for each training episode.

```
episodeCumulativeRewardVector = [];
```

Create a figure for training visualization using the `hBuildFigure` on page 5-0 helper function.

```
[trainingPlot, lineReward, lineAveReward] = hBuildFigure;
```

Custom Training loop

The algorithm for the custom training loop is as follows. For each episode:

- 1 Reset the environment.
- 2 Create buffers for storing experience information: observations, actions, and rewards.
- 3 Generate experiences until a terminal condition occurs. To do so, evaluate the policy to get actions, apply those actions to the environment, and obtain the resulting observations and rewards. Store the actions, observations, and rewards in buffers.
- 4 Collect the training data as a batch of experiences.

- 5 Compute the episode Monte Carlo return, which is the discounted future reward.
- 6 Compute the gradient of the loss function with respect to the policy representation parameters.
- 7 Update the actor representation using the computed gradients.
- 8 Update the training visualization.
- 9 Terminate training if the policy is sufficiently trained.

```
% Enable the training visualization plot.
set(trainingPlot,'Visible','on');

% Train the policy for the maximum number of episodes or until the average
% reward indicates that the policy is sufficiently trained.
for episodeCt = 1:numEpisodes

    % 1. Reset the environment at the start of the episode
    obs = reset(env);

    episodeReward = zeros(maxStepsPerEpisode,1);

    % 2. Create buffers to store experiences. The dimensions for each buffer
    % must be as follows.
    %
    % For observation buffer:
    %     numObservations x numObservationChannels x batchSize
    %
    % For action buffer:
    %     numActions x numActionChannels x batchSize
    %
    % For reward buffer:
    %     1 x batchSize
    %
    observationBuffer = zeros(numObs,1,maxStepsPerEpisode);
    actionBuffer = zeros(numAct,1,maxStepsPerEpisode);
    rewardBuffer = zeros(1,maxStepsPerEpisode);

    % 3. Generate experiences for the maximum number of steps per
    % episode or until a terminal condition is reached.
    for stepCt = 1:maxStepsPerEpisode

        % Compute an action using the policy based on the current
        % observation.
        action = getAction(actor,{obs});

        % Apply the action to the environment and obtain the resulting
        % observation and reward.
        [nextObs,reward,isdone] = step(env,action{1});

        % Store the action, observation, and reward experiences in buffers.
        observationBuffer(:,:,stepCt) = obs;
        actionBuffer(:,:,stepCt) = action{1};
        rewardBuffer(:,stepCt) = reward;

        episodeReward(stepCt) = reward;
        obs = nextObs;

        % Stop if a terminal condition is reached.
        if isdone
```

```

        break;
    end

end

% 4. Create training data. Training is performed using batch data. The
% batch size equal to the length of the episode.
batchSize = min(stepCt,maxStepsPerEpisode);
observationBatch = observationBuffer(:,:,1:batchSize);
actionBatch = actionBuffer(:,:,1:batchSize);
rewardBatch = rewardBuffer(:,1:batchSize);

% Compute the discounted future reward.
discountedReturn = zeros(1,batchSize);
for t = 1:batchSize
    G = 0;
    for k = t:batchSize
        G = G + discountFactor ^ (k-t) * rewardBatch(k);
    end
    discountedReturn(t) = G;
end

% 5. Organize data to pass to the loss function.
lossData.batchSize = batchSize;
lossData.actInfo = actInfo;
lossData.actionBatch = actionBatch;
lossData.discountedReturn = discountedReturn;

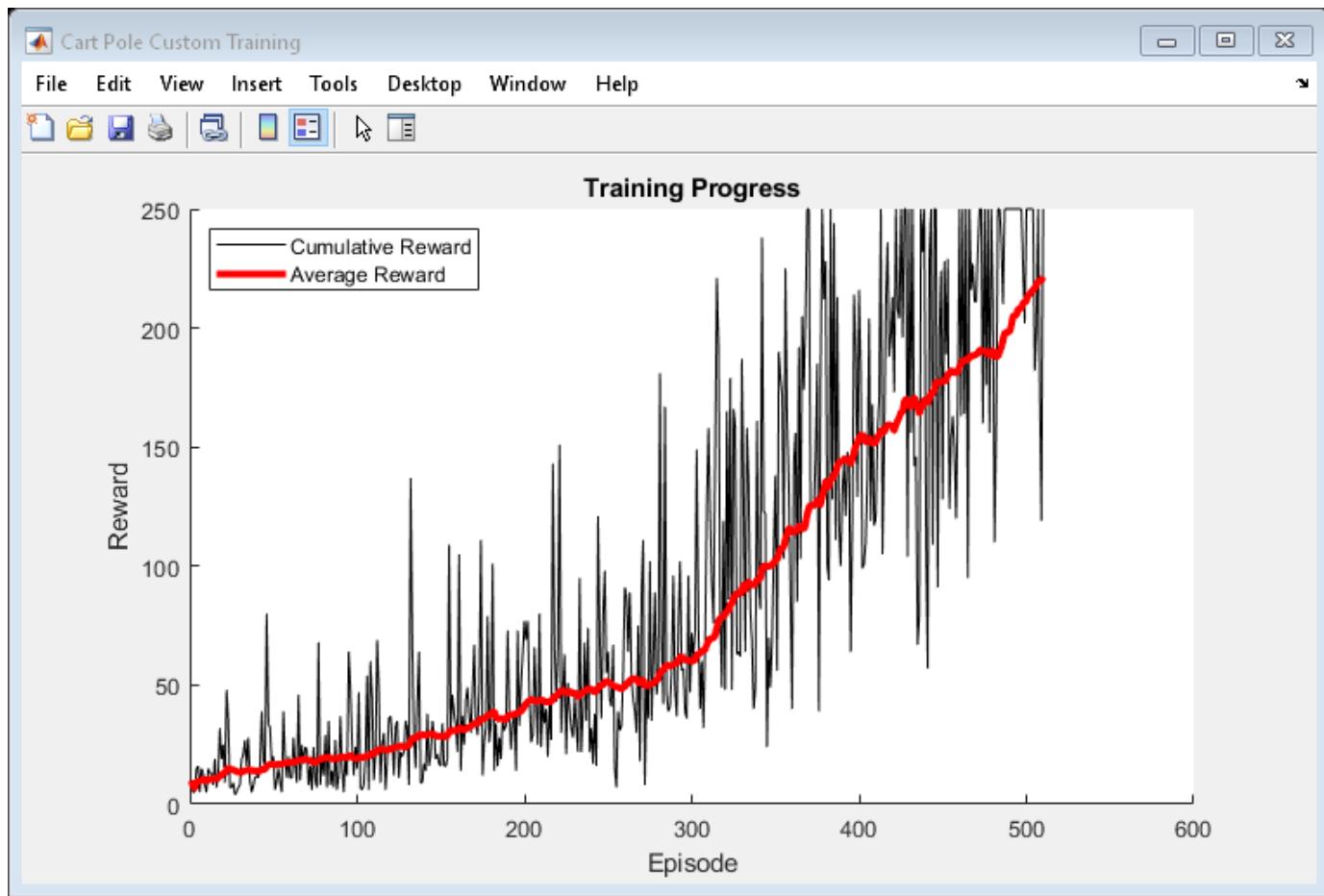
% 6. Compute the gradient of the loss with respect to the policy
% parameters.
actorGradient = gradient(actor,'loss-parameters',...
    {observationBatch},lossData);

% 7. Update the actor network using the computed gradients.
actor = optimize(actor,actorGradient);

% 8. Update the training visualization.
episodeCumulativeReward = sum(episodeReward);
episodeCumulativeRewardVector = cat(2,...
    episodeCumulativeRewardVector,episodeCumulativeReward);
movingAveReward = movmean(episodeCumulativeRewardVector,...%
    aveWindowSize,2);
addpoints(lineReward,episodeCt,episodeCumulativeReward);
addpoints(lineAveReward,episodeCt,movingAveReward(end));
drawnow;

% 9. Terminate training if the network is sufficiently trained.
if max(movingAveReward) > trainingTerminationValue
    break
end
end

```



Simulation

After training, simulate the trained policy.

Before simulation, reset the environment.

```
obs = reset(env);
```

Enable the environment visualization, which is updated each time the environment step function is called.

```
plot(env)
```

For each simulation step, perform the following actions.

- 1 Get the action by sampling from the policy using the `getAction` function.
- 2 Step the environment using the obtained action value.
- 3 Terminate if a terminal condition is reached.

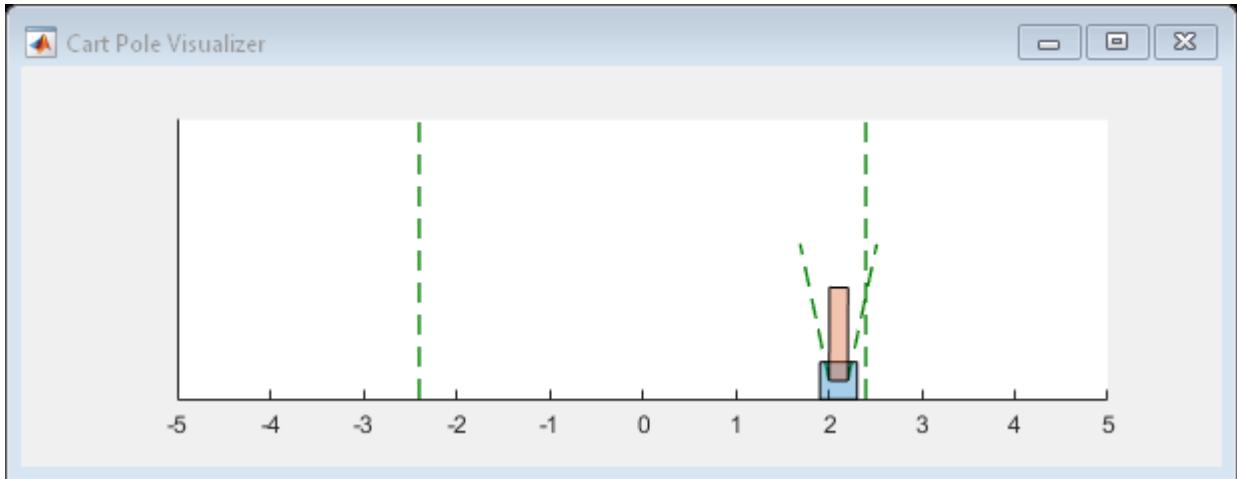
```
for stepCt = 1:maxStepsPerEpisode  
    % Select action according to trained policy  
    action = getAction(actor,{obs});
```

```
% Step the environment
[nextObs,reward,isdone] = step(env,action{1});

% Check for terminal condition
if isdone
    break
end

obs = nextObs;

end
```



Functions for Custom Training

To obtain actions and value functions for given observations from Reinforcement Learning Toolbox policy and value function representations, you can use the following functions.

- `getValue` — Obtain the estimated state value or state-action value function.
- `getAction` — Obtain the action from an actor representation based on the current observation.
- `getMaxQValue` — Obtain the estimated maximum state-action value function for a discrete Q-value representation.

If your policy or value function representation is a recurrent neural network, that is, a neural network with at least one layer that has hidden state information, the preceding functions can return the current network state. You can use the following function syntaxes to get and set the state of your representation.

- `state = getState(rep)` — Obtain the state of representation `rep`.
- `newRep = setState(oldRep,state)` — Set the state of representation `oldRep`, and return the result in `oldRep`.
- `newRep = resetState(oldRep)` — Reset all state values of `oldRep` to zero and return the result in `newRep`.

You can get and set the learnable parameters of your representation using the `getLearnableParameters` and `setLearnableParameters` function, respectively.

In addition to these functions, you can use the `setLoss`, `gradient`, `optimize`, and `syncParameters` functions to set parameters and compute gradients for your policy and value function representations.

setLoss

The policy is trained in a stochastic gradient ascent manner where the gradients of a loss function is used to update the network. For custom training, you can set the loss function using the `setLoss` function. To do so, use the following syntax.

```
newRep = setLoss(oldRep,lossFcn)
```

Here:

- `oldRep` is a policy or value function representation object.
- `lossFcn` is the name of a custom loss function or a handle to a custom loss function.
- `newRep` is equivalent to `oldRep`, except that the loss function has been added to the representation.

gradient

The `gradient` function computes the gradients of the representation loss function. You can compute several different gradients. For example, to compute the gradient of the representation outputs with respect to its inputs, use the following syntax.

```
grad = gradient(rep,"output-input",inputData)
```

Here:

- `rep` is a policy or value function representation object.
- `inputData` contains values for the input channels to the representation.
- `grad` contains the computed gradients.

For more information, at the MATLAB command line, type `help rl.representation.rlAbstractRepresentation.gradient`.

optimize

The `optimize` function updates the learnable parameters of the representation based on computed gradients. To update the parameters the gradients, use the following syntax.

```
newRep = optimize(oldRep,grad)
```

Here, `oldRep` is a policy or value function representation object and `grad` contains gradients computed using the `gradient` function. `newRep` has the same structure as `oldRep`, but its parameters are updated.

syncParameters

The `syncParameters` function updates the learnable parameters of one policy or value function representation based on those of another representation. This function is useful for updating a target actor or critic representation, as is done for DDPG agents. To synchronize parameters values between two representations, use the following syntax.

```
newTargetRep = syncParameters(oldTargetRep,sourceRep,smoothFactor)
```

Here:

- `oldTargetRep` is a policy or value function representation object with parameters θ_{old} .
- `sourceRep` is a policy or value function representation object with the same structure as `oldTargetRep`, but with parameters θ_{source} .
- `smoothFactor` is a smoothing factor (τ) for the update.
- `newTargetRep` has the same structure as `oldRep`, but its parameters are $\theta_{\text{new}} = \tau\theta_{\text{source}} + (1 - \tau)\theta_{\text{old}}$.

Loss Function

The loss function in the REINFORCE algorithm is the product of the discounted reward and the log of the policy, summed across all time steps. The discounted reward calculated in the custom training loop must be resized to make it compatible for multiplication with the policy.

```
function loss = actorLossFunction(policy, lossData)

    % Create the action indication matrix.
    batchSize = lossData.batchSize;
    Z = repmat(lossData.actInfo.Elements', 1, batchSize);
    actionIndicationMatrix = lossData.actionBatch(:, :) == Z;

    % Resize the discounted return to the size of policy.
    G = actionIndicationMatrix .* lossData.discountedReturn;
    G = reshape(G, size(policy));

    % Round any policy values less than eps to eps.
    policy(policy < eps) = eps;

    % Compute the loss.
    loss = -sum(G .* log(policy), 'all');

end
```

Helper Function

The following helper function creates a figure for training visualization.

```
function [trainingPlot, lineReward, lineAveReward] = hBuildFigure()
    plotRatio = 16/9;
    trainingPlot = figure(...%
        'Visible','off',...
        'HandleVisibility','off',...
        'NumberTitle','off',...
        'Name','Cart Pole Custom Training');
    trainingPlot.Position(3) = plotRatio * trainingPlot.Position(4);

    ax = gca(trainingPlot);

    lineReward = animatedline(ax);
    lineAveReward = animatedline(ax, 'Color','r','LineWidth',3);
    xlabel(ax, 'Episode');
    ylabel(ax, 'Reward');
    legend(ax, 'Cumulative Reward', 'Average Reward', 'Location', 'northwest')
```

```
    title(ax,'Training Progress');
end
```

See Also

Deploy Trained Policies

Deploy Trained Reinforcement Learning Policies

Once you train a reinforcement learning agent, you can generate code to deploy the optimal policy. You can generate:

- CUDA® code for deep neural network policies using GPU Coder
- C/C++ code for table, deep neural network, or linear basis function policies using MATLAB Coder

Note Code generation for deep neural network policies supports only networks with a single input layer.

For more information on training reinforcement learning agents, see “Train Reinforcement Learning Agents” on page 5-2.

Create Policy Evaluation Function

To generate code for the trained optimal policy of a reinforcement learning agent, you must first create a policy evaluation function for the agent. You can generate a policy function for an agent with any type of policy representation object:

- Value and Q tables (`rlTableRepresentation`)
- Deep neural networks (`rlLayerRepresentation`)
- Linear basis functions (`rlLinearBasisRepresentation`)

For more information on the different types of policies, see “Create Policy and Value Function Representations” on page 3-2.

To create a policy evaluation function that selects an action based on a given observation, use the `generatePolicyFunction` command. This command generates a MATLAB script, which contains the policy evaluation function, and a MAT-file, which contains the optimal policy data.

You can generate code to deploy this policy function using GPU Coder or MATLAB Coder.

Generate Code Using GPU Coder

If your trained optimal policy uses a deep neural network, you can generate CUDA code for the policy using GPU Coder. There are several required and recommended prerequisite products for generating CUDA code for deep neural networks. For more information, see “Installing Prerequisite Products” (GPU Coder) and “Setting Up the Prerequisite Products” (GPU Coder).

Not all deep neural network layers support GPU code generation. For a list of supported layers, see “Supported Networks and Layers” (GPU Coder). For more information and examples on GPU code generation, see “Deep Learning with GPU Coder” (GPU Coder).

Generate CUDA Code for Deep Neural Network Policy

As an example, generate GPU code for the policy gradient agent trained in “Train PG Agent to Balance Cart-Pole System” on page 5-14.

Load the trained agent.

```
load('MATLABCartpolePG.mat','agent')
```

Create a policy evaluation function for this agent.

```
generatePolicyFunction(agent)
```

This command creates the `evaluatePolicy.m` file, which contains the policy function, and the `agentData.mat` file, which contains the trained deep neural network actor. For a given observation, the policy function evaluates a probability for each potential action using the actor network. Then, the policy function randomly selects an action based on these probabilities.

Since the actor network for this PG agent has a single input layer and single output layer, you can generate code for this network using GPU Coder. For example, you can generate a CUDA compatible MEX function.

Configure the `codegen` function to create a CUDA compatible C++ MEX function.

```
cfg = coder.gpuConfig('mex');
cfg.TargetLang = 'C++';
cfg.DeepLearningConfig = coder.DeepLearningConfig('cudnn');
```

Set the dimensions of the policy evaluation input argument, which correspond to the observation specification dimensions for the agent. To find the observation dimensions, use the `getObservationInfo` function. In this case, the observations are in a four-element vector.

```
argstr = '{ones(4,1)}';
```

Generate code using the `codegen` function.

```
codegen('-config','cfg','evaluatePolicy',' -args',argstr,'-report');
```

This command generates the MEX function `evaluatePolicy_mex`.

Generate Code Using MATLAB Coder

You can generate C/C++ code for table, deep neural network, or linear basis function policies using MATLAB Coder.

Using MATLAB Coder, you can generate:

- C/C++ code for policies that use Q tables, value tables, or linear basis functions. For more information on general C/C++ code generation, see “Generating Code” (MATLAB Coder).
- C++ code for policies that use deep neural networks. For more information, see “Prerequisites for Deep Learning with MATLAB Coder” (MATLAB Coder) and “Deep Learning with MATLAB Coder” (MATLAB Coder).

Generate C Code for Q Table Policy

As an example, generate C code for the Q-learning agent trained in “Train Reinforcement Learning Agent in Basic Grid World” on page 1-8.

Load the trained agent.

```
load('basicGWQAgent.mat','qAgent')
```

Create a policy evaluation function for this agent.

```
generatePolicyFunction(qAgent)
```

This command creates the `evaluatePolicy.m` file, which contains the policy function, and the `agentData.mat` file, which contains the trained Q table value function. For a given observation, the policy function looks up the value function for each potential action using the Q table. Then, the policy function selects the action for which the value function is greatest.

Set the dimensions of the policy evaluation input argument, which correspond to the observation specification dimensions for the agent. To find the observation dimensions, use the `getObservationInfo` function. In this case, there is a single finite observation.

```
argstr = '{[1]};
```

Configure the `codegen` function to generate embeddable C code suitable for targeting a static library, and set the output folder to `buildFolder`.

```
cfg = coder.config('lib');
outFolder = 'buildFolder';
```

Generate C code using the `codegen` function.

```
codegen('-c', '-d', outFolder, '-config', 'cfg',...
    'evaluatePolicy', '-args', argstr, '-report');
```

See Also

`generatePolicyFunction`

More About

- “Reinforcement Learning Agents” on page 4-2
- “Train Reinforcement Learning Agents” on page 5-2