

---

# ChainerRL Documentation

*Release 0.8.0*

**Preferred Networks, Inc.**

**Feb 14, 2020**



---

# Contents

---

<b>1</b>	<b>Installation</b>	<b>3</b>
1.1	How to install ChainerRL . . . . .	3
<b>2</b>	<b>API Reference</b>	<b>5</b>
2.1	Action values . . . . .	5
2.2	Agents . . . . .	6
2.3	Distributions . . . . .	21
2.4	Experiments . . . . .	23
2.5	Explorers . . . . .	27
2.6	Links . . . . .	28
2.7	Policies . . . . .	31
2.8	Q-functions . . . . .	36
2.9	Using recurrent models . . . . .	41
2.10	Replay Buffers . . . . .	43
<b>3</b>	<b>Indices and tables</b>	<b>45</b>
	<b>Index</b>	<b>47</b>



ChainerRL is a deep reinforcement learning library that implements various state-of-the-art deep reinforcement algorithms in Python using [Chainer](#), a flexible deep learning framework.



### 1.1 How to install ChainerRL

ChainerRL is tested with 3.5.1+. For other requirements, see `requirements.txt`.

Listing 1: `requirements.txt`

```
cached-property
chainer>=4.0.0
gym>=0.9.7
numpy>=1.10.4
pillow
scipy
```

ChainerRL can be installed via PyPI,

```
pip install chainerrl
```

or through the source code:

```
git clone https://github.com/chainer/chainerrl.git
cd chainerrl
python setup.py install
```





## 2.1 Action values

### 2.1.1 Action value interfaces

**class** `chainerrl.action_value.ActionValue`  
Struct that holds state-fixed Q-functions and its subproducts.

Every operation it supports is done in a batch manner.

**evaluate\_actions** (*actions*)  
Evaluate  $Q(s,a)$  with  $a =$  given actions.

**greedy\_actions**  
Get  $\operatorname{argmax}_a Q(s,a)$ .

**max**  
Evaluate  $\max Q(s,a)$ .

**params**  
Learnable parameters of this action value.

**Returns** tuple of `chainer.Variable`

### 2.1.2 Action value implementations

**class** `chainerrl.action_value.DiscreteActionValue` (*q\_values*,  
*q\_values\_formatter*=<function  
`DiscreteActionValue.<lambda>>`)  
Q-function output for discrete action space.

**Parameters** **q\_values** (*ndarray* or *chainer.Variable*) – Array of Q values whose shape is (batchsize, n\_actions)

**class** `chainerrl.action_value.QuadraticActionValue` (*mu*, *mat*, *v*, *min\_action=None*,  
*max\_action=None*)

Q-function output for continuous action space.

See: <http://arxiv.org/abs/1603.00748>

Define a  $Q(s,a)$  with  $A(s,a)$  in a quadratic form.

$Q(s,a) = V(s,a) + A(s,a) A(s,a) = -1/2 (u - \mu(s))^T P(s) (u - \mu(s))$

#### Parameters

- **mu** (*chainer.Variable*) –  $\mu(s)$ , actions that maximize  $A(s,a)$
- **mat** (*chainer.Variable*) –  $P(s)$ , coefficient matrices of  $A(s,a)$ . It must be positive definite.
- **v** (*chainer.Variable*) –  $V(s)$ , values of  $s$
- **min\_action** (*ndarray*) – minimum action, not batched
- **max\_action** (*ndarray*) – maximum action, not batched

**class** `chainerrl.action_value.SingleActionValue` (*evaluator*, *maximizer=None*)

ActionValue that can evaluate only a single action.

## 2.2 Agents

### 2.2.1 Agent interfaces

**class** `chainerrl.agent.Agent`

Abstract agent class.

**act** (*obs*)

Select an action for evaluation.

**Returns** action

**Return type** ~object

**act\_and\_train** (*obs*, *reward*)

Select an action for training.

**Returns** action

**Return type** ~object

**get\_statistics** ()

Get statistics of the agent.

**Returns**

List of two-item tuples. The first item in a tuple is a str that represents the name of item, while the second item is a value to be recorded.

Example: [(‘average\_loss’: 0), (‘average\_value’: 1), ...]

**load** (*dirname*)

Load internal states.

**Returns** None

**save** (*dirname*)

Save internal states.

**Returns** None

**stop\_episode** ()

Prepare for a new episode.

**Returns** None

**stop\_episode\_and\_train** (*state, reward, done=False*)

Observe consequences and prepare for a new episode.

**Returns** None

## 2.2.2 Agent implementations

```
class chainerrl.agents.A2C (model, optimizer, gamma, num_processes, gpu=None, update_steps=5, phi=<function A2C.<lambda>>, pi_loss_coef=1.0, v_loss_coef=0.5, entropy_coef=0.01, use_gae=False, tau=0.95, act_deterministically=False, average_actor_loss_decay=0.999, average_entropy_decay=0.999, average_value_decay=0.999, batch_states=<function batch_states>)
```

A2C: Advantage Actor-Critic.

**A2C is a synchronous, deterministic variant of Asynchronous Advantage Actor Critic (A3C).**

See <https://arxiv.org/abs/1708.05144>

### Parameters

- **model** (*A2CModel*) – Model to train
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **gamma** (*float*) – Discount factor [0,1]
- **num\_processes** (*int*) – The number of processes
- **gpu** (*int*) – GPU device id if not None nor negative.
- **update\_steps** (*int*) – The number of update steps
- **phi** (*callable*) – Feature extractor function
- **pi\_loss\_coef** (*float*) – Weight coefficient for the loss of the policy
- **v\_loss\_coef** (*float*) – Weight coefficient for the loss of the value function
- **entropy\_coef** (*float*) – Weight coefficient for the loss of the entropy
- **use\_gae** (*bool*) – use generalized advantage estimation(GAE)
- **tau** (*float*) – gae parameter
- **average\_actor\_loss\_decay** (*float*) – Decay rate of average actor loss. Used only to record statistics.
- **average\_entropy\_decay** (*float*) – Decay rate of average entropy. Used only to record statistics.
- **average\_value\_decay** (*float*) – Decay rate of average value. Used only to record statistics.
- **act\_deterministically** (*bool*) – If set true, choose most probable actions in act method.

- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`

```
class chainerrl.agents.A3C(model, optimizer, t_max, gamma, beta=0.01, process_idx=0,
    phi=<function A3C.<lambda>>, pi_loss_coef=1.0, v_loss_coef=0.5,
    keep_loss_scale_same=False, normalize_grad_by_t_max=False,
    use_average_reward=False, average_reward_tau=0.01,
    act_deterministically=False, average_entropy_decay=0.999,
    average_value_decay=0.999, batch_states=<function batch_states>)
```

A3C: Asynchronous Advantage Actor-Critic.

See <http://arxiv.org/abs/1602.01783>

#### Parameters

- **model** (*A3CModel*) – Model to train
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **t\_max** (*int*) – The model is updated after every `t_max` local steps
- **gamma** (*float*) – Discount factor [0,1]
- **beta** (*float*) – Weight coefficient for the entropy regularization term.
- **process\_idx** (*int*) – Index of the process.
- **phi** (*callable*) – Feature extractor function
- **pi\_loss\_coef** (*float*) – Weight coefficient for the loss of the policy
- **v\_loss\_coef** (*float*) – Weight coefficient for the loss of the value function
- **act\_deterministically** (*bool*) – If set true, choose most probable actions in `act` method.
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`

```
class chainerrl.agents.ACER(model, optimizer, t_max, gamma, replay_buffer, beta=0.01,
    phi=<function ACER.<lambda>>, pi_loss_coef=1.0,
    Q_loss_coef=0.5, use_trust_region=True, trust_region_alpha=0.99,
    trust_region_delta=1, truncation_threshold=10,
    disable_online_update=False, n_times_replay=8,
    replay_start_size=10000, normalize_loss_by_steps=True,
    act_deterministically=False, use_Q_opc=False,
    average_entropy_decay=0.999, average_value_decay=0.999,
    average_kl_decay=0.999, logger=None)
```

ACER (Actor-Critic with Experience Replay).

See <http://arxiv.org/abs/1611.01224>

#### Parameters

- **model** (*ACERModel*) – Model to train. It must be a callable that accepts observations as input and return three values: action distributions (`Distribution`), Q values (`ActionValue`) and state values (`chainer.Variable`).
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **t\_max** (*int*) – The model is updated after every `t_max` local steps
- **gamma** (*float*) – Discount factor [0,1]

- **replay\_buffer** (`EpisodicReplayBuffer`) – Replay buffer to use. If set `None`, this agent won't use experience replay.
- **beta** (`float`) – Weight coefficient for the entropy regularization term.
- **phi** (`callable`) – Feature extractor function
- **pi\_loss\_coef** (`float`) – Weight coefficient for the loss of the policy
- **Q\_loss\_coef** (`float`) – Weight coefficient for the loss of the value function
- **use\_trust\_region** (`bool`) – If set true, use efficient TRPO.
- **trust\_region\_alpha** (`float`) – Decay rate of the average model used for efficient TRPO.
- **trust\_region\_delta** (`float`) – Threshold used for efficient TRPO.
- **truncation\_threshold** (`float or None`) – Threshold used to truncate larger importance weights. If set `None`, importance weights are not truncated.
- **disable\_online\_update** (`bool`) – If set true, disable online on-policy update and rely only on experience replay.
- **n\_times\_replay** (`int`) – Number of times experience replay is repeated per one time of online update.
- **replay\_start\_size** (`int`) – Experience replay is disabled if the number of transitions in the replay buffer is lower than this value.
- **normalize\_loss\_by\_steps** (`bool`) – If set true, losses are normalized by the number of steps taken to accumulate the losses
- **act\_deterministically** (`bool`) – If set true, choose most probable actions in act method.
- **use\_Q\_opc** (`bool`) – If set true, use `Q_opc`, a Q-value estimate without importance sampling, is used to compute advantage values for policy gradients. The original paper recommend to use in case of continuous action.
- **average\_entropy\_decay** (`float`) – Decay rate of average entropy. Used only to record statistics.
- **average\_value\_decay** (`float`) – Decay rate of average value. Used only to record statistics.
- **average\_kl\_decay** (`float`) – Decay rate of kl value. Used only to record statistics.

**class** `chainerrl.agents.AL` (`*args, **kwargs`)

Advantage Learning.

See: <http://arxiv.org/abs/1512.04860>.

**Parameters** **alpha** (`float`) – Weight of (persistent) advantages. Convergence is guaranteed only for alpha in [0, 1).

For other arguments, see DQN.

```
class chainerrl.agents.CategoricalDoubleDQN(q_function, optimizer, replay_buffer,
                                           gamma, explorer, gpu=None, replay_start_size=50000,
                                           minibatch_size=32, update_interval=1,
                                           target_update_interval=10000,
                                           clip_delta=True, phi=<function DQN.<lambda>>,
                                           target_update_method='hard',
                                           soft_update_tau=0.01, n_times_update=1,
                                           average_q_decay=0.999, average_loss_decay=0.99,
                                           batch_accumulator='mean',
                                           episodic_update_len=None, logger=<Logger chainerrl.agents.dqn (WARNING)>,
                                           batch_states=<function batch_states>, recurrent=False)
```

Categorical Double DQN.

```
class chainerrl.agents.CategoricalDQN(q_function, optimizer, replay_buffer, gamma,
                                      explorer, gpu=None, replay_start_size=50000,
                                      minibatch_size=32, update_interval=1, target_update_interval=10000,
                                      clip_delta=True, phi=<function DQN.<lambda>>,
                                      target_update_method='hard', soft_update_tau=0.01,
                                      n_times_update=1, average_q_decay=0.999,
                                      average_loss_decay=0.99, batch_accumulator='mean',
                                      episodic_update_len=None, logger=<Logger chainerrl.agents.dqn (WARNING)>,
                                      batch_states=<function batch_states>, recurrent=False)
```

Categorical DQN.

See <https://arxiv.org/abs/1707.06887>.

Arguments are the same as those of DQN except *q\_function* must return `DistributionalDiscreteActionValue` and *clip\_delta* is ignored.

```
class chainerrl.agents.DDPG(model, actor_optimizer, critic_optimizer, replay_buffer,
                            gamma, explorer, gpu=None, replay_start_size=50000,
                            minibatch_size=32, update_interval=1, target_update_interval=10000,
                            phi=<function DDPG.<lambda>>, target_update_method='hard',
                            soft_update_tau=0.01, n_times_update=1, average_q_decay=0.999,
                            average_loss_decay=0.99, episodic_update=False,
                            episodic_update_len=None, logger=<Logger chainerrl.agents.ddpg (WARNING)>,
                            batch_states=<function batch_states>,
                            burnin_action_func=None)
```

Deep Deterministic Policy Gradients.

This can be used as SVG(0) by specifying a Gaussian policy instead of a deterministic policy.

#### Parameters

- **model** (*DDPGModel*) – DDPG model that contains both a policy and a Q-function
- **actor\_optimizer** (*Optimizer*) – Optimizer setup with the policy
- **critic\_optimizer** (*Optimizer*) – Optimizer setup with the Q-function
- **replay\_buffer** (*ReplayBuffer*) – Replay buffer

- **gamma** (*float*) – Discount factor
- **explorer** (*Explorer*) – Explorer that specifies an exploration strategy.
- **gpu** (*int*) – GPU device id if not None nor negative.
- **replay\_start\_size** (*int*) – if the replay buffer’s size is less than `replay_start_size`, skip update
- **minibatch\_size** (*int*) – Minibatch size
- **update\_interval** (*int*) – Model update interval in step
- **target\_update\_interval** (*int*) – Target model update interval in step
- **phi** (*callable*) – Feature extractor applied to observations
- **target\_update\_method** (*str*) – ‘hard’ or ‘soft’.
- **soft\_update\_tau** (*float*) – Tau of soft target update.
- **n\_times\_update** (*int*) – Number of repetition of update
- **average\_q\_decay** (*float*) – Decay rate of average Q, only used for recording statistics
- **average\_loss\_decay** (*float*) – Decay rate of average loss, only used for recording statistics
- **batch\_accumulator** (*str*) – ‘mean’ or ‘sum’
- **episodic\_update** (*bool*) – Use full episodes for update if set True
- **episodic\_update\_len** (*int* or *None*) – Subsequences of this length are used for update if set int and `episodic_update=True`
- **logger** (*Logger*) – Logger used
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`
- **burnin\_action\_func** (*callable* or *None*) – If not None, this callable object is used to select actions before the model is updated one or more times during training.

```
class chainerrl.agents.DoubleDQN(q_function, optimizer, replay_buffer, gamma, explorer,
                                gpu=None, replay_start_size=50000, minibatch_size=32,
                                update_interval=1, target_update_interval=10000,
                                clip_delta=True, phi=<function DQN.<lambda>>,
                                target_update_method='hard', soft_update_tau=0.01,
                                n_times_update=1, average_q_decay=0.999, average_loss_decay=0.99,
                                batch_accumulator='mean',
                                episodic_update_len=None, logger=<Logger chainerrl.agents.dqn (WARNING)>,
                                batch_states=<function batch_states>, recurrent=False)
```

Double DQN.

See: <http://arxiv.org/abs/1509.06461>.

```
class chainerrl.agents.DoublePAL(*args, **kwargs)
```

```
class chainerrl.agents.DPP(*args, **kwargs)
```

Dynamic Policy Programming with softmax operator.

**Parameters** **eta** (*float*) – Positive constant.

For other arguments, see DQN.

```
class chainerrl.agents.DQN(q_function, optimizer, replay_buffer, gamma, explorer,  
                        gpu=None, replay_start_size=50000, minibatch_size=32, up-  
date_interval=1, target_update_interval=10000, clip_delta=True,  
phi=<function DQN.<lambda>>, target_update_method='hard',  
soft_update_tau=0.01, n_times_update=1, average_q_decay=0.999,  
average_loss_decay=0.99, batch_accumulator='mean',  
episodic_update_len=None, logger=<Logger chainerrl.agents.dqn  
(WARNING)>, batch_states=<function batch_states>, recur-  
rent=False)
```

Deep Q-Network algorithm.

#### Parameters

- **q\_function** (*StateQFunction*) – Q-function
- **optimizer** (*Optimizer*) – Optimizer that is already setup
- **replay\_buffer** (*ReplayBuffer*) – Replay buffer
- **gamma** (*float*) – Discount factor
- **explorer** (*Explorer*) – Explorer that specifies an exploration strategy.
- **gpu** (*int*) – GPU device id if not None nor negative.
- **replay\_start\_size** (*int*) – if the replay buffer’s size is less than `replay_start_size`, skip update
- **minibatch\_size** (*int*) – Minibatch size
- **update\_interval** (*int*) – Model update interval in step
- **target\_update\_interval** (*int*) – Target model update interval in step
- **clip\_delta** (*bool*) – Clip delta if set True
- **phi** (*callable*) – Feature extractor applied to observations
- **target\_update\_method** (*str*) – ‘hard’ or ‘soft’.
- **soft\_update\_tau** (*float*) – Tau of soft target update.
- **n\_times\_update** (*int*) – Number of repetition of update
- **average\_q\_decay** (*float*) – Decay rate of average Q, only used for recording statistics
- **average\_loss\_decay** (*float*) – Decay rate of average loss, only used for recording statistics
- **batch\_accumulator** (*str*) – ‘mean’ or ‘sum’
- **episodic\_update\_len** (*int or None*) – Subsequences of this length are used for update if set int and `episodic_update=True`
- **logger** (*Logger*) – Logger used
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`
- **recurrent** (*bool*) – If set to True, `model` is assumed to implement `chainerrl.links.StatelessRecurrent` and is updated in a recurrent manner.

```
class chainerrl.agents.IQN(*args, **kwargs)
```

Implicit Quantile Networks.

See <https://arxiv.org/abs/1806.06923>.



**Parameters**

- **quantile\_thresholds\_N** (*int*) – Number of quantile thresholds used in quantile regression.
- **quantile\_thresholds\_N\_prime** (*int*) – Number of quantile thresholds used to sample from the return distribution at the next state.
- **quantile\_thresholds\_K** (*int*) – Number of quantile thresholds used to compute greedy actions.
- **act\_deterministically** (*bool*) – IQN’s action selection is by default stochastic as it samples quantile thresholds every time it acts, even for evaluation. If this option is set to True, it uses equally spaced quantile thresholds instead of randomly sampled ones for evaluation, making its action selection deterministic.

For other arguments, see `chainerrl.agents.DQN`.

```
class chainerrl.agents.NSQ(q_function, optimizer, t_max, gamma, i_target, explorer,
                          phi=<function NSQ.<lambda>>, average_q_decay=0.999,
                          logger=<Logger chainerrl.agents.nsq (WARNING)>,
                          batch_states=<function batch_states>)
```

Asynchronous N-step Q-Learning.

See <http://arxiv.org/abs/1602.01783>

**Parameters**

- **q\_function** (*A3CModel*) – Model to train
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **t\_max** (*int*) – The model is updated after every `t_max` local steps
- **gamma** (*float*) – Discount factor [0,1]
- **i\_target** (*intn*) – The target model is updated after every `i_target` global steps
- **explorer** (*Explorer*) – Explorer to use in training
- **phi** (*callable*) – Feature extractor function
- **average\_q\_decay** (*float*) – Decay rate of average Q, only used for recording statistics
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`

```
class chainerrl.agents.PAL(*args, **kwargs)
    Persistent Advantage Learning.
```

See: <http://arxiv.org/abs/1512.04860>.

**Parameters** **alpha** (*float*) – Weight of (persistent) advantages. Convergence is guaranteed only for alpha in [0, 1).

For other arguments, see DQN.

```
class chainerrl.agents.PCL(model, optimizer, replay_buffer=None, t_max=None,
                           gamma=0.99, tau=0.01, phi=<function PCL.<lambda>>,
                           pi_loss_coef=1.0, v_loss_coef=0.5, rollout_len=10, batch-
                           size=1, disable_online_update=False, n_times_replay=1,
                           replay_start_size=100, normalize_loss_by_steps=True,
                           act_deterministically=False, average_loss_decay=0.999, av-
                           erage_entropy_decay=0.999, average_value_decay=0.999, ex-
                           plorer=None, logger=None, batch_states=<function batch_states>,
                           backprop_future_values=True, train_async=False)
```

PCL (Path Consistency Learning).

Not only the batch PCL algorithm proposed in the paper but also its asynchronous variant is implemented.

See <https://arxiv.org/abs/1702.08892>

### Parameters

- **model** (*chainer.Link*) – Model to train. It must be a callable that accepts a batch of observations as input and return two values:
  - action distributions (*Distribution*)
  - state values (*chainer.Variable*)
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **t\_max** (*int* or *None*) – The model is updated after every t\_max local steps. If set None, the model is updated after every episode.
- **gamma** (*float*) – Discount factor [0,1]
- **tau** (*float*) – Weight coefficient for the entropy regularization term.
- **phi** (*callable*) – Feature extractor function
- **pi\_loss\_coef** (*float*) – Weight coefficient for the loss of the policy
- **v\_loss\_coef** (*float*) – Weight coefficient for the loss of the value function
- **rollout\_len** (*int*) – Number of rollout steps
- **batchsize** (*int*) – Number of episodes or sub-trajectories used for an update. The total number of transitions used will be (batchsize x t\_max).
- **disable\_online\_update** (*bool*) – If set true, disable online on-policy update and rely only on experience replay.
- **n\_times\_replay** (*int*) – Number of times experience replay is repeated per one time of online update.
- **replay\_start\_size** (*int*) – Experience replay is disabled if the number of transitions in the replay buffer is lower than this value.
- **normalize\_loss\_by\_steps** (*bool*) – If set true, losses are normalized by the number of steps taken to accumulate the losses
- **act\_deterministically** (*bool*) – If set true, choose most probable actions in act method.
- **average\_loss\_decay** (*float*) – Decay rate of average loss. Used only to record statistics.
- **average\_entropy\_decay** (*float*) – Decay rate of average entropy. Used only to record statistics.

- **average\_value\_decay** (*float*) – Decay rate of average value. Used only to record statistics.
- **explorer** (*Explorer* or *None*) – If not *None*, this explorer is used for selecting actions.
- **logger** (*None* or *Logger*) – Logger to be used
- **batch\_states** (*callable*) – Method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`
- **backprop\_future\_values** (*bool*) – If set *True*, value gradients are computed not only wrt  $V(s_t)$  but also  $V(s_{t+d})$ .
- **train\_async** (*bool*) – If set *True*, use a process-local model to compute gradients and update the globally shared model.

```
class chainerrl.agents.PGT(model, actor_optimizer, critic_optimizer, replay_buffer, gamma,
                          explorer, beta=0.01, act_deterministically=False, gpu=-1, replay_start_size=50000,
                          minibatch_size=32, update_interval=1, target_update_interval=10000,
                          phi=<function PGT.<lambda>>, target_update_method='hard', soft_update_tau=0.01,
                          n_times_update=1, average_q_decay=0.999, average_loss_decay=0.99,
                          logger=<Logger chainerrl.agents.pgt (WARNING)>, batch_states=<function batch_states>)
```

Policy Gradient Theorem with an approximate policy and a Q-function.

This agent is almost the same with DDPG except that it uses the likelihood ratio gradient estimation instead of value gradients.

#### Parameters

- **model** (*chainer.Chain*) – Chain that contains both a policy and a Q-function
- **actor\_optimizer** (*Optimizer*) – Optimizer setup with the policy
- **critic\_optimizer** (*Optimizer*) – Optimizer setup with the Q-function
- **replay\_buffer** (*ReplayBuffer*) – Replay buffer
- **gamma** (*float*) – Discount factor
- **explorer** (*Explorer*) – Explorer that specifies an exploration strategy.
- **gpu** (*int*) – GPU device id. -1 for CPU.
- **replay\_start\_size** (*int*) – if the replay buffer's size is less than `replay_start_size`, skip update
- **minibatch\_size** (*int*) – Minibatch size
- **update\_interval** (*int*) – Model update interval in step
- **target\_update\_interval** (*int*) – Target model update interval in step
- **phi** (*callable*) – Feature extractor applied to observations
- **target\_update\_method** (*str*) – 'hard' or 'soft'.
- **soft\_update\_tau** (*float*) – Tau of soft target update.
- **n\_times\_update** (*int*) – Number of repetition of update
- **average\_q\_decay** (*float*) – Decay rate of average Q, only used for recording statistics
- **average\_loss\_decay** (*float*) – Decay rate of average loss, only used for recording statistics

- **batch\_accumulator** (*str*) – ‘mean’ or ‘sum’
- **logger** (*Logger*) – Logger used
- **beta** (*float*) – Coefficient for entropy regularization
- **act\_deterministically** (*bool*) – Act deterministically by selecting most probable actions in test time
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`

```
class chainerrl.agents.PPO(model, optimizer, obs_normalizer=None, gpu=None, gamma=0.99,
                          lambd=0.95, phi=<function PPO.<lambda>>, value_func_coef=1.0,
                          entropy_coef=0.01, update_interval=2048, minibatch_size=64,
                          epochs=10, clip_eps=0.2, clip_eps_vf=None, standardize_advantages=True,
                          batch_states=<function batch_states>, recurrent=False,
                          max_recurrent_sequence_len=None, act_deterministically=False,
                          value_stats_window=1000, entropy_stats_window=1000,
                          value_loss_stats_window=100, policy_loss_stats_window=100)
```

Proximal Policy Optimization

See <https://arxiv.org/abs/1707.06347>

#### Parameters

- **model** (*A3CModel*) – Model to train. Recurrent models are not supported. state  $s \rightarrow (\pi(s, _), v(s))$
- **optimizer** (*chainer.Optimizer*) – Optimizer used to train the model
- **gpu** (*int*) – GPU device id if not None nor negative
- **gamma** (*float*) – Discount factor [0, 1]
- **lambd** (*float*) – Lambda-return factor [0, 1]
- **phi** (*callable*) – Feature extractor function
- **value\_func\_coef** (*float*) – Weight coefficient for loss of value function (0, inf)
- **entropy\_coef** (*float*) – Weight coefficient for entropy bonus [0, inf]
- **update\_interval** (*int*) – Model update interval in step
- **minibatch\_size** (*int*) – Minibatch size
- **epochs** (*int*) – Training epochs in an update
- **clip\_eps** (*float*) – Epsilon for pessimistic clipping of likelihood ratio to update policy
- **clip\_eps\_vf** (*float*) – Epsilon for pessimistic clipping of value to update value function. If it is None, value function is not clipped on updates.
- **standardize\_advantages** (*bool*) – Use standardized advantages on updates
- **recurrent** (*bool*) – If set to True, *model* is assumed to implement `chainerrl.links.StatelessRecurrent` and update in a recurrent manner.
- **max\_recurrent\_sequence\_len** (*int*) – Maximum length of consecutive sequences of transitions in a minibatch for updating the model. This value is used only when *recurrent* is True. A smaller value will encourage a minibatch to contain more and shorter sequences.
- **act\_deterministically** (*bool*) – If set to True, choose most probable actions in the act method instead of sampling from distributions.

- **value\_stats\_window** (*int*) – Window size used to compute statistics of value predictions.
- **entropy\_stats\_window** (*int*) – Window size used to compute statistics of entropy of action distributions.
- **value\_loss\_stats\_window** (*int*) – Window size used to compute statistics of loss values regarding the value function.
- **policy\_loss\_stats\_window** (*int*) – Window size used to compute statistics of loss values regarding the policy.

#### Statistics:

**average\_value:** Average of value predictions on non-terminal states. It's updated on **(batch\_)act\_and\_train**.

**average\_entropy:** Average of entropy of action distributions on non-terminal states. It's updated on **(batch\_)act\_and\_train**.

**average\_value\_loss:** Average of losses regarding the value function. It's updated after the model is updated.

**average\_policy\_loss:** Average of losses regarding the policy. It's updated after the model is updated.

**n\_updates:** Number of model updates so far. **explained\_variance:** Explained variance computed from the last batch.

```
class chainerrl.agents.REINFORCE (model, optimizer, beta=0, phi=<function
    REINFORCE.<lambda>>, batchsize=1,
    act_deterministically=False, average_entropy_decay=0.999,
    backward_separately=False, batch_states=<function
    batch_states>, logger=None)
```

William's episodic REINFORCE.

#### Parameters

- **model** (*Policy*) – Model to train. It must be a callable that accepts observations as input and return action distributions (*Distribution*).
- **optimizer** (*chainer.Optimizer*) – optimizer used to train the model
- **beta** (*float*) – Weight coefficient for the entropy regularization term.
- **normalize\_loss\_by\_steps** (*bool*) – If set true, losses are normalized by the number of steps taken to accumulate the losses
- **act\_deterministically** (*bool*) – If set true, choose most probable actions in act method.
- **batchsize** (*int*) – Number of episodes used for each update
- **backward\_separately** (*bool*) – If set true, call backward separately for each episode and accumulate only gradients.
- **average\_entropy\_decay** (*float*) – Decay rate of average entropy. Used only to record statistics.
- **batch\_states** (*callable*) – Method which makes a batch of observations. default is *chainerrl.misc.batch\_states*
- **logger** (*logging.Logger*) – Logger to be used.

```
class chainerrl.agents.ResidualDQN(*args, **kwargs)
```

DQN that allows maxQ also backpropagate gradients.

```
class chainerrl.agents.SARSA(q_function, optimizer, replay_buffer, gamma, explorer,  
gpu=None, replay_start_size=50000, minibatch_size=32,  
update_interval=1, target_update_interval=10000,  
clip_delta=True, phi=<function DQN.<lambda>>,  
target_update_method='hard', soft_update_tau=0.01,  
n_times_update=1, average_q_decay=0.999, aver-  
age_loss_decay=0.99, batch_accumulator='mean',  
episodic_update_len=None, logger=<Logger chainerrl.agents.dqn  
(WARNING)>, batch_states=<function batch_states>, recur-  
rent=False)
```

Off-policy SARSA.

This agent learns the Q-function of a behavior policy defined via the given explorer, instead of learning the Q-function of the optimal policy.

```
class chainerrl.agents.SoftActorCritic(policy, q_func1, q_func2, policy_optimizer,  
q_func1_optimizer, q_func2_optimizer,  
replay_buffer, gamma, gpu=None, re-  
play_start_size=10000, minibatch_size=100,  
update_interval=1, phi=<function SoftActor-  
Critic.<lambda>>, soft_update_tau=0.005, log-  
ger=<Logger chainerrl.agents.soft_actor_critic  
(WARNING)>, batch_states=<function  
batch_states>, burnin_action_func=None, ini-  
tial_temperature=1.0, entropy_target=None,  
temperature_optimizer=None,  
act_deterministically=True)
```

Soft Actor-Critic (SAC).

See <https://arxiv.org/abs/1812.05905>

### Parameters

- **policy** (*Policy*) – Policy.
- **q\_func1** (*Link*) – First Q-function that takes state-action pairs as input and outputs predicted Q-values.
- **q\_func2** (*Link*) – Second Q-function that takes state-action pairs as input and outputs predicted Q-values.
- **policy\_optimizer** (*Optimizer*) – Optimizer setup with the policy
- **q\_func1\_optimizer** (*Optimizer*) – Optimizer setup with the first Q-function.
- **q\_func2\_optimizer** (*Optimizer*) – Optimizer setup with the second Q-function.
- **replay\_buffer** (*ReplayBuffer*) – Replay buffer
- **gamma** (*float*) – Discount factor
- **gpu** (*int*) – GPU device id if not None nor negative.
- **replay\_start\_size** (*int*) – if the replay buffer's size is less than `replay_start_size`, skip update
- **minibatch\_size** (*int*) – Minibatch size
- **update\_interval** (*int*) – Model update interval in step

- **phi** (*callable*) – Feature extractor applied to observations
- **soft\_update\_tau** (*float*) – Tau of soft target update.
- **logger** (*Logger*) – Logger used
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`
- **burnin\_action\_func** (*callable or None*) – If not None, this callable object is used to select actions before the model is updated one or more times during training.
- **initial\_temperature** (*float*) – Initial temperature value. If `entropy_target` is set to None, the temperature is fixed to it.
- **entropy\_target** (*float or None*) – If set to a float, the temperature is adjusted during training to match the policy’s entropy to it.
- **temperature\_optimizer** (*Optimizer or None*) – Optimizer used to optimize the temperature. If set to None, Adam with default hyperparameters is used.
- **act\_deterministically** (*bool*) – If set to True, choose most probable actions in the act method instead of sampling from distributions.

```
class chainerrl.agents.TD3(policy, q_func1, q_func2, policy_optimizer, q_func1_optimizer,
                          q_func2_optimizer, replay_buffer, gamma, explorer,
                          gpu=None, replay_start_size=10000, minibatch_size=100,
                          update_interval=1, phi=<function TD3.<lambda>>,
                          soft_update_tau=0.005, n_times_update=1, logger=<Logger
chainerrl.agents.td3 (WARNING)>, batch_states=<function
batch_states>, burnin_action_func=None, pol-
icy_update_delay=2, target_policy_smoothing_func=<function
default_target_policy_smoothing_func>)
```

Twin Delayed Deep Deterministic Policy Gradients (TD3).

See <http://arxiv.org/abs/1802.09477>

### Parameters

- **policy** (*Policy*) – Policy.
- **q\_func1** (*Link*) – First Q-function that takes state-action pairs as input and outputs predicted Q-values.
- **q\_func2** (*Link*) – Second Q-function that takes state-action pairs as input and outputs predicted Q-values.
- **policy\_optimizer** (*Optimizer*) – Optimizer setup with the policy
- **q\_func1\_optimizer** (*Optimizer*) – Optimizer setup with the first Q-function.
- **q\_func2\_optimizer** (*Optimizer*) – Optimizer setup with the second Q-function.
- **replay\_buffer** (*ReplayBuffer*) – Replay buffer
- **gamma** (*float*) – Discount factor
- **explorer** (*Explorer*) – Explorer that specifies an exploration strategy.
- **gpu** (*int*) – GPU device id if not None nor negative.
- **replay\_start\_size** (*int*) – if the replay buffer’s size is less than `replay_start_size`, skip update
- **minibatch\_size** (*int*) – Minibatch size



- **update\_interval** (*int*) – Model update interval in step
- **phi** (*callable*) – Feature extractor applied to observations
- **soft\_update\_tau** (*float*) – Tau of soft target update.
- **logger** (*Logger*) – Logger used
- **batch\_states** (*callable*) – method which makes a batch of observations. default is `chainerrl.misc.batch_states.batch_states`
- **burnin\_action\_func** (*callable or None*) – If not None, this callable object is used to select actions before the model is updated one or more times during training.
- **policy\_update\_delay** (*int*) – Delay of policy updates. Policy is updated once in `policy_update_delay` times of Q-function updates.
- **target\_policy\_smoothing\_func** (*callable*) – Callable that takes a batch of actions as input and outputs a noisy version of it. It is used for target policy smoothing when computing target Q-values.

```
class chainerrl.agents.TRPO (policy, vf, vf_optimizer, obs_normalizer=None, gamma=0.99,
                             lambda=0.95, phi=<function TRPO.<lambda>>, entropy_coef=0.01,
                             update_interval=2048, max_kl=0.01, vf_epochs=3,
                             vf_batch_size=64, standardize_advantages=True,
                             batch_states=<function batch_states>, recurrent=False,
                             max_recurrent_sequence_len=None, line_search_max_backtrack=10,
                             conjugate_gradient_max_iter=10, conjugate_gradient_damping=0.01,
                             act_deterministically=False, value_stats_window=1000,
                             entropy_stats_window=1000, kl_stats_window=100,
                             policy_step_size_stats_window=100, logger=<Logger chainerrl.agents.trpo (WARNING)>)
```

Trust Region Policy Optimization.

A given stochastic policy is optimized by the TRPO algorithm. A given value function is also trained to predict by the TD(lambda) algorithm and used for Generalized Advantage Estimation (GAE).

Since the policy is optimized via the conjugate gradient method and line search while the value function is optimized via SGD, these two models should be separate.

Since TRPO requires second-order derivatives to compute Hessian-vector products, Chainer v3.0.0 or newer is required. In addition, your policy must contain only functions that support second-order derivatives.

See <https://arxiv.org/abs/1502.05477> for TRPO. See <https://arxiv.org/abs/1506.02438> for GAE.

#### Parameters

- **policy** (*Policy*) – Stochastic policy. Its forward computation must contain only functions that support second-order derivatives. Recurrent models are not supported.
- **vf** (*ValueFunction*) – Value function. Recurrent models are not supported.
- **vf\_optimizer** (*chainer.Optimizer*) – Optimizer for the value function.
- **obs\_normalizer** (*chainerrl.links.EmpiricalNormalization or None*) – If set to `chainerrl.links.EmpiricalNormalization`, it is used to normalize observations based on the empirical mean and standard deviation of observations. These statistics are updated after computing advantages and target values and before updating the policy and the value function.
- **gamma** (*float*) – Discount factor [0, 1]
- **lambda** (*float*) – Lambda-return factor [0, 1]



- **phi** (*callable*) – Feature extractor function
- **entropy\_coef** (*float*) – Weight coefficient for entropy bonus [0, inf)
- **update\_interval** (*int*) – Interval steps of TRPO iterations. Every after this amount of steps, this agent updates the policy and the value function using data from these steps.
- **vf\_epochs** (*int*) – Number of epochs for which the value function is trained on each TRPO iteration.
- **vf\_batch\_size** (*int*) – Batch size of SGD for the value function.
- **standardize\_advantages** (*bool*) – Use standardized advantages on updates
- **line\_search\_max\_backtrack** (*int*) – Maximum number of backtracking in line search to tune step sizes of policy updates.
- **conjugate\_gradient\_max\_iter** (*int*) – Maximum number of iterations in the conjugate gradient method.
- **conjugate\_gradient\_damping** (*float*) – Damping factor used in the conjugate gradient method.
- **act\_deterministically** (*bool*) – If set to True, choose most probable actions in the act method instead of sampling from distributions.
- **value\_stats\_window** (*int*) – Window size used to compute statistics of value predictions.
- **entropy\_stats\_window** (*int*) – Window size used to compute statistics of entropy of action distributions.
- **kl\_stats\_window** (*int*) – Window size used to compute statistics of KL divergence between old and new policies.
- **policy\_step\_size\_stats\_window** (*int*) – Window size used to compute statistics of step sizes of policy updates.

#### Statistics:

**average\_value:** Average of value predictions on non-terminal states. It's updated before the value function is updated.

**average\_entropy:** Average of entropy of action distributions on non-terminal states. It's updated on `act_and_train`.

**average\_kl:** Average of KL divergence between old and new policies. It's updated after the policy is updated.

**average\_policy\_step\_size:** Average of step sizes of policy updates. It's updated after the policy is updated.

## 2.3 Distributions

### 2.3.1 Distribution interfaces

**class** `chainerrl.distribution.Distribution`

Batch of distributions of data.

**copy** (*x*)

Copy a distribution unchained from the computation graph.

**Returns** Distribution

**entropy**

Entropy of distributions.

**Returns** chainer.Variable

**kl**

Compute KL divergence  $D_{KL}(P|Q)$ .

**Parameters** **distrib** (*Distribution*) – Distribution Q.

**Returns** chainer.Variable

**log\_prob** (*x*)

Compute  $\log p(x)$ .

**Returns** chainer.Variable

**most\_probable**

Most probable data points.

**Returns** chainer.Variable

**params**

Learnable parameters of this distribution.

**Returns** tuple of chainer.Variable

**prob** (*x*)

Compute  $p(x)$ .

**Returns** chainer.Variable

**sample** ()

Sample from distributions.

**Returns** chainer.Variable

**sample\_with\_log\_prob** ()

Do *sample* and *log\_prob* at the same time.

This can be more efficient than calling *sample* and *log\_prob* separately.

**Returns** Samples. chainer.Variable: Log probability of the samples.

**Return type** chainer.Variable

## 2.3.2 Distribution implementations

**class** chainerrl.distribution.**GaussianDistribution** (*mean, var*)  
Gaussian distribution.

**class** chainerrl.distribution.**SoftmaxDistribution** (*logits, beta=1.0, min\_prob=0.0*)  
Softmax distribution.

**Parameters**

- **logits** (*ndarray or chainer.Variable*) – Logits for softmax distribution.
- **beta** (*float*) – inverse of the temperature parameter of softmax distribution
- **min\_prob** (*float*) – minimum probability across all labels

**class** `chainerrl.distribution.MellowmaxDistribution` (*values*, *omega*=8.0)  
Maximum entropy mellowmax distribution.

See: <http://arxiv.org/abs/1612.05628>

**Parameters** *values* (*ndarray* or *chainer.Variable*) – Values to apply mellowmax.

**class** `chainerrl.distribution.ContinuousDeterministicDistribution` (*x*)  
Continuous deterministic distribution.

This distribution is supposed to be used in continuous deterministic policies.

## 2.4 Experiments

### 2.4.1 Collecting demonstrations

`chainerrl.experiments.collect_demonstrations` (*agent*, *env*, *steps*, *episodes*, *outdir*,  
*max\_episode\_len*=None, *logger*=None)

Collects demonstrations from an agent and writes them to a file.

#### Parameters

- **agent** – Agent from which demonstrations are collected.
- **env** – Environment in which the agent produces demonstrations.
- **steps** (*int*) – Number of total time steps to collect demonstrations for.
- **episodes** (*int*) – Number of total episodes to collect demonstrations for.
- **outdir** (*str*) – Path to the directory to output demonstrations.
- **max\_episode\_len** (*int*) – Maximum episode length.
- **logger** (*logging.Logger*) – Logger used in this function.

### 2.4.2 Training and evaluation

`chainerrl.experiments.train_agent_async` (*outdir*, *processes*, *make\_env*, *profile*=False,  
*steps*=80000000, *eval\_interval*=1000000,  
*eval\_n\_steps*=None, *eval\_n\_episodes*=10,  
*max\_episode\_len*=None, *step\_offset*=0,  
*successful\_score*=None, *agent*=None,  
*make\_agent*=None, *global\_step\_hooks*=(),  
*save\_best\_so\_far\_agent*=True, *logger*=None)

Train agent asynchronously using multiprocessing.

Either *agent* or *make\_agent* must be specified.

#### Parameters

- **outdir** (*str*) – Path to the directory to output things.
- **processes** (*int*) – Number of processes.
- **make\_env** (*callable*) – (process\_idx, test) -> Environment.
- **profile** (*bool*) – Profile if set True.
- **steps** (*int*) – Number of global time steps for training.

- **eval\_interval** (*int*) – Interval of evaluation. If set to None, the agent will not be evaluated at all.
- **eval\_n\_steps** (*int*) – Number of eval timesteps at each eval phase
- **eval\_n\_episodes** (*int*) – Number of eval episodes at each eval phase
- **max\_episode\_len** (*int*) – Maximum episode length.
- **step\_offset** (*int*) – Time step from which training starts.
- **successful\_score** (*float*) – Finish training if the mean score is greater or equal to this value if not None
- **agent** (*Agent*) – Agent to train.
- **make\_agent** (*callable*) – (process\_idx) -> Agent
- **global\_step\_hooks** (*Sequence*) – Sequence of callable objects that accepts (env, agent, step) as arguments. They are called every global step. See `chainerrl.experiments.hooks`.
- **save\_best\_so\_far\_agent** (*bool*) – If set to True, after each evaluation, if the score (= mean return of evaluation episodes) exceeds the best-so-far score, the current agent is saved.
- **logger** (*logging.Logger*) – Logger used in this function.

**Returns** Trained agent.

`chainerrl.experiments.train_agent_batch` (*agent, env, steps, outdir, checkpoint\_freq=None, log\_interval=None, max\_episode\_len=None, eval\_interval=None, step\_offset=0, evaluator=None, successful\_score=None, step\_hooks=(), return\_window\_size=100, logger=None*)

Train an agent in a batch environment.

#### Parameters

- **agent** – Agent to train.
- **env** – Environment to train the agent against.
- **steps** (*int*) – Number of total time steps for training.
- **eval\_interval** (*int*) – Interval of evaluation.
- **outdir** (*str*) – Path to the directory to output things.
- **checkpoint\_freq** (*int*) – frequency at which agents are stored.
- **log\_interval** (*int*) – Interval of logging.
- **max\_episode\_len** (*int*) – Maximum episode length.
- **step\_offset** (*int*) – Time step from which training starts.
- **return\_window\_size** (*int*) – Number of training episodes used to estimate the average returns of the current agent.
- **successful\_score** (*float*) – Finish training if the mean score is greater or equal to this value if not None
- **step\_hooks** (*Sequence*) – Sequence of callable objects that accepts (env, agent, step) as arguments. They are called every step. See `chainerrl.experiments.hooks`.
- **logger** (*logging.Logger*) – Logger used in this function.

```
chainerrl.experiments.train_agent_batch_with_evaluation(agent, env, steps,
                                                       eval_n_steps,
                                                       eval_n_episodes,
                                                       eval_interval, outdir,
                                                       checkpoint_freq=None,
                                                       max_episode_len=None,
                                                       step_offset=0,
                                                       eval_max_episode_len=None,
                                                       return_window_size=100,
                                                       eval_env=None,
                                                       log_interval=None,
                                                       successful_score=None,
                                                       step_hooks=(),
                                                       save_best_so_far_agent=True,
                                                       logger=None)
```

Train an agent while regularly evaluating it.

#### Parameters

- **agent** – Agent to train.
- **env** – Environment train the agent against.
- **steps** (*int*) – Number of total time steps for training.
- **eval\_n\_steps** (*int*) – Number of timesteps at each evaluation phase.
- **eval\_n\_runs** (*int*) – Number of runs for each time of evaluation.
- **eval\_interval** (*int*) – Interval of evaluation.
- **outdir** (*str*) – Path to the directory to output things.
- **log\_interval** (*int*) – Interval of logging.
- **checkpoint\_freq** (*int*) – frequency with which to store networks
- **max\_episode\_len** (*int*) – Maximum episode length.
- **step\_offset** (*int*) – Time step from which training starts.
- **return\_window\_size** (*int*) – Number of training episodes used to estimate the average returns of the current agent.
- **eval\_max\_episode\_len** (*int or None*) – Maximum episode length of evaluation runs. If set to None, max\_episode\_len is used instead.
- **eval\_env** – Environment used for evaluation.
- **successful\_score** (*float*) – Finish training if the mean score is greater or equal to this value if not None
- **step\_hooks** (*Sequence*) – Sequence of callable objects that accepts (env, agent, step) as arguments. They are called every step. See chainerrl.experiments.hooks.
- **save\_best\_so\_far\_agent** (*bool*) – If set to True, after each evaluation, if the score (= mean return of evaluation episodes) exceeds the best-so-far score, the current agent is saved.
- **logger** (*logging.Logger*) – Logger used in this function.

```
chainerrl.experiments.train_agent_with_evaluation(agent, env, steps, eval_n_steps,
                                                eval_n_episodes, eval_interval,
                                                outdir, checkpoint_freq=None,
                                                train_max_episode_len=None,
                                                step_offset=0,
                                                eval_max_episode_len=None,
                                                eval_env=None, successful_score=None,
                                                step_hooks=(),
                                                save_best_so_far_agent=True,
                                                logger=None)
```

Train an agent while periodically evaluating it.

#### Parameters

- **agent** – A `chainerrl.agent.Agent`
- **env** – Environment train the agent against.
- **steps** (*int*) – Total number of timesteps for training.
- **eval\_n\_steps** (*int*) – Number of timesteps at each evaluation phase.
- **eval\_n\_episodes** (*int*) – Number of episodes at each evaluation phase.
- **eval\_interval** (*int*) – Interval of evaluation.
- **outdir** (*str*) – Path to the directory to output data.
- **checkpoint\_freq** (*int*) – frequency at which agents are stored.
- **train\_max\_episode\_len** (*int*) – Maximum episode length during training.
- **step\_offset** (*int*) – Time step from which training starts.
- **eval\_max\_episode\_len** (*int or None*) – Maximum episode length of evaluation runs. If `None`, `train_max_episode_len` is used instead.
- **eval\_env** – Environment used for evaluation.
- **successful\_score** (*float*) – Finish training if the mean score is greater than or equal to this value if not `None`
- **step\_hooks** (*Sequence*) – Sequence of callable objects that accepts (`env`, `agent`, `step`) as arguments. They are called every step. See `chainerrl.experiments.hooks`.
- **save\_best\_so\_far\_agent** (*bool*) – If set to `True`, after each evaluation phase, if the score (= mean return of evaluation episodes) exceeds the best-so-far score, the current agent is saved.
- **logger** (*logging.Logger*) – Logger used in this function.

### 2.4.3 Training hooks

**class** `chainerrl.experiments.StepHook`

Hook function that will be called in training.

This class is for clarifying the interface required for Hook functions. You don't need to inherit this class to define your own hooks. Any callable that accepts (`env`, `agent`, `step`) as arguments can be used as a hook.

**class** `chainerrl.experiments.LinearInterpolationHook` (*total\_steps*, *start\_value*,  
*stop\_value*, *setter*)

Hook that will set a linearly interpolated value.

You can use this hook to decay the learning rate by using a setter function as follows:

```
def lr_setter(env, agent, value):
    agent.optimizer.lr = value

hook = LinearInterpolationHook(10 ** 6, 1e-3, 0, lr_setter)
```

### Parameters

- **total\_steps** (*int*) – Number of total steps.
- **start\_value** (*float*) – Start value.
- **stop\_value** (*float*) – Stop value.
- **setter** (*callable*) – (env, agent, value) -> None

## 2.5 Explorers

### 2.5.1 Explorer interfaces

**class** `chainerrl.explorer.Explorer`

Abstract explorer.

**select\_action** (*t*, *greedy\_action\_func*, *action\_value=None*)

Select an action.

### Parameters

- **t** – current time step
- **greedy\_action\_func** – function with no argument that returns an action
- **action\_value** (`ActionValue`) – ActionValue object

### 2.5.2 Explorer implementations

**class** `chainerrl.explorers.AdditiveGaussian` (*scale*, *low=None*, *high=None*)

Additive Gaussian noise to actions.

Each action must be `numpy.ndarray`.

### Parameters

- **scale** (*float or array\_like of floats*) – Scale parameter.
- **low** (*float, array\_like of floats, or None*) – Lower bound of action space used to clip an action after adding a noise. If set to None, clipping is not performed on lower edge.
- **high** (*float, array\_like of floats, or None*) – Higher bound of action space used to clip an action after adding a noise. If set to None, clipping is not performed on upper edge.

**class** `chainerrl.explorers.AdditiveOU` (*mu=0.0*, *theta=0.15*, *sigma=0.3*, *start\_with\_mu=False*,  
*logger=<Logger chainerrl.explorers.additive\_ou (WARNING)>*)

Additive Ornstein-Uhlenbeck process.

Used in <https://arxiv.org/abs/1509.02971> for exploration.

#### Parameters

- **mu** (*float*) – Mean of the OU process
- **theta** (*float*) – Friction to pull towards the mean
- **sigma** (*float or ndarray*) – Scale of noise
- **start\_with\_mu** (*bool*) – Start the process without noise

**class** `chainerrl.explorers.Boltzmann` (*T=1.0*)  
Boltzmann exploration.

**Parameters** **T** (*float*) – Temperature of Boltzmann distribution.

**class** `chainerrl.explorers.ConstantEpsilonGreedy` (*epsilon*, *random\_action\_func*,  
*logger=<Logger chainerrl.explorers.epsilon\_greedy (WARNING)>*)

Epsilon-greedy with constant epsilon.

#### Parameters

- **epsilon** – epsilon used
- **random\_action\_func** – function with no argument that returns action
- **logger** – logger used

**class** `chainerrl.explorers.LinearDecayEpsilonGreedy` (*start\_epsilon*, *end\_epsilon*, *decay\_steps*, *random\_action\_func*,  
*logger=<Logger chainerrl.explorers.epsilon\_greedy (WARNING)>*)

Epsilon-greedy with linearly decayed epsilon

#### Parameters

- **start\_epsilon** – max value of epsilon
- **end\_epsilon** – min value of epsilon
- **decay\_steps** – how many steps it takes for epsilon to decay
- **random\_action\_func** – function with no argument that returns action
- **logger** – logger used

**class** `chainerrl.explorers.Greedy`  
No exploration

## 2.6 Links

### 2.6.1 Link interfaces

**class** `chainerrl.links.StatelessRecurrent`  
Stateless recurrent link interface.

This class defines the interface of a recurrent link ChainerRL can handle.



In most cases, you can just use ChainerRL's existing containers like `chainerrl.links.StatelessRecurrentChainList`, `chainerrl.links.StatelessRecurrentSequential`, and `chainerrl.links.StatelessRecurrentBranched` to define a recurrent link. You can use Chainer's recurrent links such as `L.NStepLSTM` inside the containers.

To write your own recurrent link, you need to implement the interface.

**concatenate\_recurrent\_states** (*split\_recurrent\_states*)

Concatenate recurrent states into a batch.

This method can be used to make a batched recurrent state from separate recurrent states obtained via the `get_recurrent_state_at` method.

**Parameters** `split_recurrent_states` (*object*) – Recurrent states to concatenate.

**Returns** Batched recurrent\_state.

**Return type** `object`

**get\_recurrent\_state\_at** (*recurrent\_state, indices*)

Get a recurrent state at given indices.

This method can be used to save a recurrent state so that you can reuse it when you replay past sequences.

**Parameters** `indices` (*int or array-like of ints*) – Which recurrent state to get.

**Returns** Recurrent state of given indices.

**Return type** `object`

**mask\_recurrent\_state\_at** (*recurrent\_state, indices*)

Return a recurrent state masked at given indices.

This method can be used to initialize a recurrent state only for a certain sequence, not all the sequences.

**Parameters**

- `recurrent_state` (*object*) – Batched recurrent state.
- `indices` (*int or array-like of ints*) – Which recurrent state to mask.

**Returns** New batched recurrent state.

**Return type** `object`

**n\_step\_forward** (*x, recurrent\_state*)

Multi-step batch forward computation.

This method sequentially applies layers as `chainer.Sequential` does.

**Parameters**

- `x` (*list*) – Input sequences. Each sequence should be a variable whose first axis corresponds to time or a tuple of such variables.
- `recurrent_state` (*object*) – Batched recurrent state. If set to `None`, it is initialized.
- `output_mode` (*str*) – If set to `'concat'`, the output value is concatenated into a single large batch, which can be suitable for loss computation. If set to `'split'`, the output value is a list of output sequences.

**Returns**

**Output sequences.** See the description of the `output_mode` argument.

`object`: New batched recurrent state.

**Return type** `object`

## 2.6.2 Link implementations

**class** `chainerrl.links.Branched` (\*links)

Link that calls forward functions of child links in parallel.

When either the `__call__` method of this link are called, all the arguments are forwarded to each child link's `__call__` method.

The returned values from the child links are returned as a tuple.

**Parameters** \*links – Child links. Each link should be callable.

**class** `chainerrl.links.EmpiricalNormalization` (shape, batch\_axis=0, eps=0.01, dtype=<class 'numpy.float32'>, until=None, clip\_threshold=None)

Normalize mean and variance of values based on empirical values.

**Parameters**

- **shape** (*int* or *tuple of int*) – Shape of input values except batch axis.
- **batch\_axis** (*int*) – Batch axis.
- **eps** (*float*) – Small value for stability.
- **dtype** (*dtype*) – Dtype of input values.
- **until** (*int* or *None*) – If this arg is specified, the link learns input values until the sum of batch sizes exceeds it.

**class** `chainerrl.links.FactorizedNoisyLinear` (mu\_link, sigma\_scale=0.4)

Linear layer in Factorized Noisy Network

**Parameters**

- **mu\_link** (*L.Linear*) – Linear link that computes mean of output.
- **sigma\_scale** (*float*) – The hyperparameter `sigma_0` in the original paper. Scaling factor of the initial weights of noise-scaling parameters.

**class** `chainerrl.links.MLP` (in\_size, out\_size, hidden\_sizes, nonlinearity=<function relu>, last\_wscales=1)

Multi-Layer Perceptron

**class** `chainerrl.links.MLPBN` (in\_size, out\_size, hidden\_sizes, normalize\_input=True, normalize\_output=False, nonlinearity=<function relu>, last\_wscales=1)

Multi-Layer Perceptron with Batch Normalization.

**Parameters**

- **in\_size** (*int*) – Input size.
- **out\_size** (*int*) – Output size.
- **hidden\_sizes** (*list of ints*) – Sizes of hidden channels.
- **normalize\_input** (*bool*) – If set to True, Batch Normalization is applied to inputs.
- **normalize\_output** (*bool*) – If set to True, Batch Normalization is applied to outputs.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported.
- **last\_wscales** (*float*) – Scale of weight initialization of the last layer.

**class** `chainerrl.links.NIPSDQNHead` (*n\_input\_channels=4, n\_output\_channels=256, activation=<function relu>, bias=0.1*)  
 DQN's head (NIPS workshop version)

**class** `chainerrl.links.NatureDQNHead` (*n\_input\_channels=4, n\_output\_channels=512, activation=<function relu>, bias=0.1*)  
 DQN's head (Nature version)

**class** `chainerrl.links.Sequence` (*\*layers*)  
 Sequential callable Link that consists of other Links.

**class** `chainerrl.links.StatelessRecurrentBranched` (*\*links*)  
 Stateless recurrent parallel link.

This is a recurrent analog to `chainerrl.links.Branched`. It bundles multiple links that implements *StatelessRecurrent*.

**Parameters** *\*links* – Child links. Each link should be recurrent and callable.

**class** `chainerrl.links.StatelessRecurrentChainList` (*\*links*)  
 ChainList that automatically handles recurrent states.

This link extends `chainer.ChainList` by adding the *recurrent\_children* property that returns all the recurrent child links and implementing recurrent state manipulation methods required for the *StatelessRecurrent* interface.

A recurrent state for this link is defined as a tuple of recurrent states of child recurrent links.

**class** `chainerrl.links.StatelessRecurrentSequential` (*\*layers*)  
 Sequential model that can contain stateless recurrent links.

This link a stateless recurrent analog to `chainer.Sequential`. It supports the stateless recurrent interface by automatically detecting recurrent links and handles recurrent states properly.

For non-recurrent layers (non-link callables or non-recurrent callable links), this link automatically concatenates the input to the layers for efficient computation.

**Parameters** *\*layers* – Callable objects.

## 2.6.3 Link utility functions

`chainerrl.links.to_factorized_noisy` (*link, \*args, \*\*kwargs*)

Add noisiness to components of given link

Currently this function supports `L.Linear` (with and without bias)

## 2.7 Policies

### 2.7.1 Policy interfaces

**class** `chainerrl.policy.Policy`

Abstract policy.

`__call__` (*state*)

Evaluate a policy.

**Returns** Distribution of actions

## 2.7.2 Policy implementations

**class** `chainerrl.policies.ContinuousDeterministicPolicy` (*model*, *model\_call=None*,  
*action\_filter=None*)

Continuous deterministic policy.

### Parameters

- **model** (*chainer.Link*) – Link that is callable and outputs action values.
- **model\_call** (*callable or None*) – Callable used instead of `model.__call__` if not `None`
- **action\_filter** (*callable or None*) – Callable applied to the outputs of the model if not `None`

**class** `chainerrl.policies.FCDeterministicPolicy` (*n\_input\_channels*, *n\_hidden\_layers*,  
*n\_hidden\_channels*, *action\_size*,  
*min\_action=None*, *max\_action=None*,  
*bound\_action=True*, *nonlinearity=<function relu>*, *last\_wscales=1.0*)

Fully-connected deterministic policy.

### Parameters

- **n\_input\_channels** (*int*) – Number of input channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **action\_size** (*int*) – Size of actions.
- **min\_action** (*ndarray or None*) – Minimum action. Used only if `bound_action` is set to `True`.
- **min\_action** – Minimum action. Used only if `bound_action` is set to `True`.
- **bound\_action** (*bool*) – If set to `True`, actions are bounded to `[min_action, max_action]` by `tanh`.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a `Variable` as an argument and return a `Variable` with the same shape. Nonlinearities with learnable parameters such as `PReLU` are not supported. It is not used if `n_hidden_layers` is zero.
- **last\_wscales** (*float*) – Scale of weight initialization of the last layer.

**class** `chainerrl.policies.FCBNDeterministicPolicy` (*n\_input\_channels*, *n\_hidden\_layers*,  
*n\_hidden\_channels*, *action\_size*, *min\_action=None*,  
*max\_action=None*,  
*bound\_action=True*, *normalize\_input=True*, *nonlinearity=<function relu>*,  
*last\_wscales=1.0*)

Fully-connected deterministic policy with Batch Normalization.

### Parameters

- **n\_input\_channels** (*int*) – Number of input channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.

- **action\_size** (*int*) – Size of actions.
- **min\_action** (*ndarray or None*) – Minimum action. Used only if `bound_action` is set to `True`.
- **min\_action** – Minimum action. Used only if `bound_action` is set to `True`.
- **bound\_action** (*bool*) – If set to `True`, actions are bounded to `[min_action, max_action]` by `tanh`.
- **normalize\_input** (*bool*) – If set to `True`, Batch Normalization is applied to inputs as well as hidden activations.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a `Variable` as an argument and return a `Variable` with the same shape. Nonlinearities with learnable parameters such as `PReLU` are not supported. It is not used if `n_hidden_layers` is zero.
- **last\_wscale** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.policies.FCLSTMDeterministicPolicy (n_input_channels,
                                                    n_hidden_layers,
                                                    n_hidden_channels,          ac
                                                    action_size,          tion_size,          min_action=None,
                                                    max_action=None,
                                                    bound_action=True,          non
                                                    linearity=<function relu>,          relu>,
                                                    last_wscale=1.0)
```

Fully-connected deterministic policy with LSTM.

#### Parameters

- **n\_input\_channels** (*int*) – Number of input channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **action\_size** (*int*) – Size of actions.
- **min\_action** (*ndarray or None*) – Minimum action. Used only if `bound_action` is set to `True`.
- **min\_action** – Minimum action. Used only if `bound_action` is set to `True`.
- **bound\_action** (*bool*) – If set to `True`, actions are bounded to `[min_action, max_action]` by `tanh`.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a `Variable` as an argument and return a `Variable` with the same shape. Nonlinearities with learnable parameters such as `PReLU` are not supported.
- **last\_wscale** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.policies.FCGaussianPolicy (n_input_channels,          action_size,
                                           n_hidden_layers=0, n_hidden_channels=None,
                                           min_action=None,          max_action=None,
                                           bound_mean=False,          var_type='spherical',
                                           nonlinearity=<function relu>, mean_wscale=1,
                                           var_wscale=1, var_bias=0, min_var=0)
```

Gaussian policy that consists of fully-connected layers.

This model has two output layers: the mean layer and the variance layer. The mean of the Gaussian is computed as follows:

Let  $y$  as the output of the mean layer. If `bound_mean=False`:

$\text{mean} = y$  (if `bound_mean=False`)

**If `bound_mean=True`:**  $\text{mean} = \text{min\_action} + \tanh(y) * (\text{max\_action} - \text{min\_action}) / 2$

**The variance of the Gaussian is computed as follows:** Let  $y$  as the output of the variance layer.  $\text{variance} = \text{softplus}(y) + \text{min\_var}$

### Parameters

- **`n_input_channels`** (*int*) – Number of input channels.
- **`action_size`** (*int*) – Number of dimensions of the action space.
- **`n_hidden_layers`** (*int*) – Number of hidden layers.
- **`n_hidden_channels`** (*int*) – Number of hidden channels.
- **`min_action`** (*ndarray*) – Minimum action. Used only when `bound_mean=True`.
- **`max_action`** (*ndarray*) – Maximum action. Used only when `bound_mean=True`.
- **`var_type`** (*str*) – Type of parameterization of variance. It must be ‘spherical’ or ‘diagonal’.
- **`nonlinearity`** (*callable*) – Nonlinearity placed between layers.
- **`mean_wscale`** (*float*) – Scale of weight initialization of the mean layer.
- **`var_wscale`** (*float*) – Scale of weight initialization of the variance layer.
- **`var_bias`** (*float*) – The initial value of the bias parameter for the variance layer.
- **`min_var`** (*float*) – Minimum value of the variance.

```
class chainerrl.policies.FCGaussianPolicyWithStateIndependentCovariance(n_input_channels,  
                                                                    ac-  
                                                                    tion_size,  
                                                                    n_hidden_layers=0,  
                                                                    n_hidden_channels=None,  
                                                                    min_action=None,  
                                                                    max_action=None,  
                                                                    bound_mean=False,  
                                                                    var_type='spherical',  
                                                                    non-  
                                                                    lin-  
                                                                    ear-  
                                                                    ity=<function  
                                                                    relu>,  
                                                                    mean_wscale=1,  
                                                                    var_func=<function  
                                                                    soft-  
                                                                    plus>,  
                                                                    var_param_init=0)
```

Gaussian policy that consists of FC layers with parametrized covariance.

This model has one output layers: the mean layer. The mean of the Gaussian is computed in the same way as FCGaussianPolicy.

### Parameters

- **n\_input\_channels** (*int*) – Number of input channels.
- **action\_size** (*int*) – Number of dimensions of the action space.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **min\_action** (*ndarray*) – Minimum action. Used only when bound\_mean=True.
- **max\_action** (*ndarray*) – Maximum action. Used only when bound\_mean=True.
- **var\_type** (*str*) – Type of parameterization of variance. It must be ‘spherical’ or ‘diagonal’.
- **nonlinearity** (*callable*) – Nonlinearity placed between layers.
- **mean\_wscale** (*float*) – Scale of weight initialization of the mean layer.
- **var\_func** (*callable*) – Callable that computes the variance from the var parameter. It should always return positive values.
- **var\_param\_init** (*float*) – Initial value the var parameter.

```
class chainerrl.policies.FCGaussianPolicyWithFixedCovariance (n_input_channels,
                                                         action_size, var,
                                                         n_hidden_layers=0,
                                                         n_hidden_channels=None,
                                                         min_action=None,
                                                         max_action=None,
                                                         bound_mean=False,
                                                         nonlinear-
                                                         ity=<function
                                                         relu>,
                                                         mean_wscale=1)
```

Gaussian policy that consists of FC layers with fixed covariance.

This model has one output layers: the mean layer. The mean of the Gaussian is computed in the same way as FCGaussianPolicy. The variance of the Gaussian must be specified as an argument.

#### Parameters

- **n\_input\_channels** (*int*) – Number of input channels.
- **action\_size** (*int*) – Number of dimensions of the action space.
- **var** (*float or ndarray*) – Variance of the Gaussian distribution.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **min\_action** (*ndarray*) – Minimum action. Used only when bound\_mean=True.
- **max\_action** (*ndarray*) – Maximum action. Used only when bound\_mean=True.
- **nonlinearity** (*callable*) – Nonlinearity placed between layers.
- **mean\_wscale** (*float*) – Scale of weight initialization of the mean layer.

```
class chainerrl.policies.GaussianHeadWithStateIndependentCovariance (action_size,
                                                                    var_type='spherical',
                                                                    var_func=<function
                                                                    soft-
                                                                    plus>,
                                                                    var_param_init=0)
```

Gaussian head with state-independent learned covariance.

This link is intended to be attached to a neural network that outputs the mean of a Gaussian policy. The only learnable parameter this link has determines the variance in a state-independent way.

State-independent parameterization of the variance of a Gaussian policy is often used with PPO and TRPO, e.g., in <https://arxiv.org/abs/1709.06560>.

#### Parameters

- **action\_size** (*int*) – Number of dimensions of the action space.
- **var\_type** (*str*) – Type of parameterization of variance. It must be ‘spherical’ or ‘diagonal’.
- **var\_func** (*callable*) – Callable that computes the variance from the var parameter. It should always return positive values.
- **var\_param\_init** (*float*) – Initial value the var parameter.

**class** `chainerrl.policies.MellowmaxPolicy` (*model*, *omega=1.0*)  
Mellowmax policy.

See: <http://arxiv.org/abs/1612.05628>

#### Parameters

- **model** (*chainer.Link*) – Link that is callable and outputs action values.
- **omega** (*float*) – Parameter of the mellowmax function.

**class** `chainerrl.policies.SoftmaxPolicy` (*model*, *beta=1.0*, *min\_prob=0.0*)  
Softmax policy that uses Boltzmann distributions.

#### Parameters

- **model** (*chainer.Link*) – Link that is callable and outputs action values.
- **beta** (*float*) – Parameter of Boltzmann distributions.

**class** `chainerrl.policies.FCSoftmaxPolicy` (*n\_input\_channels*, *n\_actions*,  
*n\_hidden\_layers=0*, *n\_hidden\_channels=None*,  
*beta=1.0*, *nonlinearity=<function relu>*,  
*last\_wscales=1.0*, *min\_prob=0.0*)

Softmax policy that consists of FC layers and rectifiers

## 2.8 Q-functions

### 2.8.1 Q-function interfaces

**class** `chainerrl.q_function.StateQFunction`  
Abstract Q-function with state input.

`__call__` (*x*)  
Evaluates Q-function

**Parameters** **x** (*ndarray*) – state input

**Returns** An instance of `ActionValue` that allows to calculate the Q-values for state *x* and every possible action

**class** `chainerrl.q_function.StateActionQFunction`  
Abstract Q-function with state and action input.



`__call__(x, a)`  
Evaluates Q-function

**Parameters**

- **x** (*ndarray*) – state input
- **a** (*ndarray*) – action input

**Returns** Q-value for state x and action a

## 2.8.2 Q-function implementations

**class** `chainerrl.q_functions.DuelingDQN` (*n\_actions*, *n\_input\_channels=4*, *activation=<function relu>*, *bias=0.1*)  
Dueling Q-Network

See: <http://arxiv.org/abs/1511.06581>

**class** `chainerrl.q_functions.DistributionalDuelingDQN` (*n\_actions*, *n\_atoms*, *v\_min*, *v\_max*, *n\_input\_channels=4*, *activation=<function relu>*, *bias=0.1*)  
Distributional dueling fully-connected Q-function with discrete actions.

**class** `chainerrl.q_functions.SingleModelStateQFunctionWithDiscreteAction` (*model*)  
Q-function with discrete actions.

**Parameters** **model** (*chainer.Link*) – Link that is callable and outputs action values.

**class** `chainerrl.q_functions.FCStateQFunctionWithDiscreteAction` (*ndim\_obs*, *n\_actions*, *n\_hidden\_channels*, *n\_hidden\_layers*, *nonlinearity=<function relu>*, *last\_wscales=1.0*)  
Fully-connected state-input Q-function with discrete actions.

**Parameters**

- **n\_dim\_obs** – number of dimensions of observation space
- **n\_actions** (*int*) – Number of actions in action space.
- **n\_hidden\_channels** – number of hidden channels
- **n\_hidden\_layers** – number of hidden layers
- **nonlinearity** (*callable*) – Nonlinearity applied after each hidden layer.
- **last\_wscales** (*float*) – Weight scale of the last layer.

**class** `chainerrl.q_functions.DistributionalSingleModelStateQFunctionWithDiscreteAction` (*model*, *z\_values*)  
Distributional Q-function with discrete actions.

**Parameters**

- **model** (*chainer.Link*) – Link that is callable and outputs atoms for each action.
- **z\_values** (*ndarray*) – Returns represented by atoms. Its shape must be (n\_atoms,).

```
class chainerrl.q_functions.DistributionalFCStateQFunctionWithDiscreteAction (ndim_obs,
                                                                              n_actions,
                                                                              n_atoms,
                                                                              v_min,
                                                                              v_max,
                                                                              n_hidden_channels,
                                                                              n_hidden_layers,
                                                                              non-
                                                                              lin-
                                                                              ear-
                                                                              ity=<function
                                                                              relu>,
                                                                              last_wscales=1.0)
```

Distributional fully-connected Q-function with discrete actions.

#### Parameters

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_actions** (*int*) – Number of actions in action space.
- **n\_atoms** (*int*) – Number of atoms of return distribution.
- **v\_min** (*float*) – Minimum value this model can approximate.
- **v\_max** (*float*) – Maximum value this model can approximate.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **nonlinearity** (*callable*) – Nonlinearity applied after each hidden layer.
- **last\_wscales** (*float*) – Weight scale of the last layer.

```
class chainerrl.q_functions.FCLSTMStateQFunction (n_dim_obs, n_dim_action,
                                                  n_hidden_channels,
                                                  n_hidden_layers)
```

Fully-connected + LSTM state-input discrete Q-function.

#### Parameters

- **n\_dim\_obs** – number of dimensions of observation space
- **n\_dim\_action** – number of dimensions of action space
- **n\_hidden\_channels** – number of hidden channels before LSTM
- **n\_hidden\_layers** – number of hidden layers before LSTM

```
class chainerrl.q_functions.FCQuadraticStateQFunction (n_input_channels,
                                                       n_dim_action,
                                                       n_hidden_channels,
                                                       n_hidden_layers,
                                                       action_space,
                                                       scale_mu=True)
```

Fully-connected state-input continuous Q-function.

See: <https://arxiv.org/abs/1603.00748>

#### Parameters

- **n\_input\_channels** – number of input channels
- **n\_dim\_action** – number of dimensions of action space

- **n\_hidden\_channels** – number of hidden channels
- **n\_hidden\_layers** – number of hidden layers
- **action\_space** – action\_space
- **scale\_mu** (*bool*) – scale mu by applying tanh if True

```
class chainerrl.q_functions.FCBNQuadraticStateQFunction(n_input_channels,
                                                    n_dim_action,
                                                    n_hidden_channels,
                                                    n_hidden_layers,
                                                    action_space,
                                                    scale_mu=True,      nor-
                                                    malize_input=True)
```

Fully-connected + BN state-input continuous Q-function.

See: <https://arxiv.org/abs/1603.00748>

#### Parameters

- **n\_input\_channels** – number of input channels
- **n\_dim\_action** – number of dimensions of action space
- **n\_hidden\_channels** – number of hidden channels
- **n\_hidden\_layers** – number of hidden layers
- **action\_space** – action\_space
- **scale\_mu** (*bool*) – scale mu by applying tanh if True
- **normalize\_input** (*bool*) – If set to True, Batch Normalization is applied to the observations

```
class chainerrl.q_functions.SingleModelStateActionQFunction(model)
```

Q-function with discrete actions.

**Parameters** *model* (*chainer.Link*) – Link that is callable and outputs action values.

```
class chainerrl.q_functions.FCSAQFunction(n_dim_obs, n_dim_action, n_hidden_channels,
                                         n_hidden_layers,      nonlinearity=<function
                                         relu>, last_wscales=1.0)
```

Fully-connected (s,a)-input Q-function.

#### Parameters

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_dim\_action** (*int*) – Number of dimensions of action space.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported. It is not used if n\_hidden\_layers is zero.
- **last\_wscales** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.q_functions.FCLSTMSAQFunction(n_dim_obs,      n_dim_action,
                                             n_hidden_channels, n_hidden_layers,
                                             nonlinearity=<function      relu>,
                                             last_wscales=1.0)
```

Fully-connected + LSTM (s,a)-input Q-function.

**Parameters**

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_dim\_action** (*int*) – Number of dimensions of action space.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported.
- **last\_wscale** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.q_functions.FCBNSAQFunction(n_dim_obs, n_dim_action,  
                                           n_hidden_channels, n_hidden_layers, normalize_input=True, nonlinearity=<function relu>, last_wscale=1.0)
```

Fully-connected + BN (s,a)-input Q-function.

**Parameters**

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_dim\_action** (*int*) – Number of dimensions of action space.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers.
- **normalize\_input** (*bool*) – If set to True, Batch Normalization is applied to both observations and actions.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported. It is not used if n\_hidden\_layers is zero.
- **last\_wscale** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.q_functions.FCBNLateActionSAQFunction(n_dim_obs, n_dim_action,  
                                                    n_hidden_channels,  
                                                    n_hidden_layers, normalize_input=True, nonlinearity=<function relu>,  
                                                    last_wscale=1.0)
```

Fully-connected + BN (s,a)-input Q-function with late action input.

Actions are not included until the second hidden layer and not normalized. This architecture is used in the DDPG paper: <http://arxiv.org/abs/1509.02971>

**Parameters**

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_dim\_action** (*int*) – Number of dimensions of action space.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers. It must be greater than or equal to 1.
- **normalize\_input** (*bool*) – If set to True, Batch Normalization is applied

- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported.
- **last\_wscales** (*float*) – Scale of weight initialization of the last layer.

```
class chainerrl.q_functions.FCLateActionSAQFunction(n_dim_obs,      n_dim_action,
                                                n_hidden_channels,
                                                n_hidden_layers,      nonlinearity=<function
                                                relu>,
                                                last_wscales=1.0)
```

Fully-connected (s,a)-input Q-function with late action input.

Actions are not included until the second hidden layer and not normalized. This architecture is used in the DDPG paper: <http://arxiv.org/abs/1509.02971>

#### Parameters

- **n\_dim\_obs** (*int*) – Number of dimensions of observation space.
- **n\_dim\_action** (*int*) – Number of dimensions of action space.
- **n\_hidden\_channels** (*int*) – Number of hidden channels.
- **n\_hidden\_layers** (*int*) – Number of hidden layers. It must be greater than or equal to 1.
- **nonlinearity** (*callable*) – Nonlinearity between layers. It must accept a Variable as an argument and return a Variable with the same shape. Nonlinearities with learnable parameters such as PReLU are not supported.
- **last\_wscales** (*float*) – Scale of weight initialization of the last layer.

## 2.9 Using recurrent models

### 2.9.1 Recurrent model interface

```
class chainerrl.recurrent.Recurrent
```

Interface of recurrent and stateful models.

This is an interface of recurrent and stateful models. ChainerRL supports recurrent neural network models as stateful models that implement this interface.

To implement this interface, you need to implement three abstract methods of it: `get_state`, `set_state` and `reset_state`.

```
get_state ()
```

Get the current state of this model.

**Returns** Any object that represents a state of this model.

```
reset_state ()
```

Reset the state of this model to the initial state.

For typical RL models, this method is expected to be called before every episode.

```
set_state (state)
```

Overwrite the state of this model with a given state.

**Parameters** **state** (*object*) – Any object that represents a state of this model.

**update\_state** (\*args, \*\*kwargs)

Update this model's state as if self.\_\_call\_\_ is called.

Unlike \_\_call\_\_, stateless objects may do nothing.

## 2.9.2 Utilities

chainerrl.recurrent.**state\_kept** (link)

Keeps the previous state of a given link.

This is a context manager that saves the current state of the link before entering the context, and then restores the saved state after escaping the context.

This will just ignore non-Recurrent links.

```
# Suppose the link is in a state A
assert link.get_state() is A

with state_kept(link):
    # The link is still in a state A
    assert link.get_state() is A

    # After evaluating the link, it may be in a different state
    y1 = link(x1)
    assert link.get_state() is not A

# After escaping from the context, the link is in a state A again
# because of the context manager
assert link.get_state() is A
```

chainerrl.recurrent.**state\_reset** (link)

Reset the state while keeping the previous state of a given link.

This is a context manager that saves the current state of the link and reset it to the initial state before entering the context, and then restores the saved state after escaping the context.

This will just ignore non-Recurrent links.

```
# Suppose the link is in a non-initial state A
assert link.get_state() is A

with state_reset(link):
    # The link's state has been reset to the initial state
    assert link.get_state() is InitialState

    # After evaluating the link, it may be in a different state
    y1 = link(x1)
    assert link.get_state() is not InitialState

# After escaping from the context, the link is in a state A again
# because of the context manager
assert link.get_state() is A
```

## 2.10 Replay Buffers

### 2.10.1 ReplayBuffer interfaces

**class** `chainerrl.replay_buffer.ReplayBuffer` (*capacity=None, num\_steps=1*)  
Experience Replay Buffer

As described in <https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf>.

#### Parameters

- **capacity** (*int*) – capacity in terms of number of transitions
- **num\_steps** (*int*) – Number of timesteps per stored transition (for N-step updates)

**append** (*state, action, reward, next\_state=None, next\_action=None, is\_state\_terminal=False, env\_id=0, \*\*kwargs*)

Append a transition to this replay buffer.

#### Parameters

- **state** – *s<sub>t</sub>*
- **action** – *a<sub>t</sub>*
- **reward** – *r<sub>t</sub>*
- **next\_state** – *s<sub>{t+1}</sub>* (can be None if terminal)
- **next\_action** – *a<sub>{t+1}</sub>* (can be None for off-policy algorithms)
- **is\_state\_terminal** (*bool*) –
- **env\_id** (*object*) – Object that is unique to each env. It indicates which env a given transition came from in multi-env training.
- **\*\*kwargs** – Any other information to store.

**load** (*filename*)

Load the content of the buffer from a file.

**Parameters** **filename** (*str*) – Path to a file.

**sample** (*num\_experiences*)

Sample *n* unique transitions from this replay buffer.

**Parameters** **n** (*int*) – Number of transitions to sample.

**Returns** Sequence of *n* sampled transitions.

**save** (*filename*)

Save the content of the buffer to a file.

**Parameters** **filename** (*str*) – Path to a file.

### 2.10.2 ReplayBuffer implementations

**class** `chainerrl.replay_buffer.EpisodicReplayBuffer` (*capacity=None*)

**class** `chainerrl.replay_buffer.ReplayBuffer` (*capacity=None, num\_steps=1*)  
Experience Replay Buffer

As described in <https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf>.

### Parameters

- **capacity** (*int*) – capacity in terms of number of transitions
- **num\_steps** (*int*) – Number of timesteps per stored transition (for N-step updates)

```
class chainerrl.replay_buffer.PrioritizedReplayBuffer (capacity=None, alpha=0.6, beta0=0.4, betasteps=200000.0, eps=0.01, normalize_by_max=True, error_min=0, error_max=1, num_steps=1)
```

### Stochastic Prioritization

<https://arxiv.org/pdf/1511.05952.pdf> Section 3.3 proportional prioritization

### Parameters

- **capacity** (*int*) – capacity in terms of number of transitions
- **alpha** (*float*) – Exponent of errors to compute probabilities to sample
- **beta0** (*float*) – Initial value of beta
- **betasteps** (*int*) – Steps to anneal beta to 1
- **eps** (*float*) – To revisit a step after its error becomes near zero
- **normalize\_by\_max** (*bool*) – Method to normalize weights. 'batch' or True (default): divide by the maximum weight in the sampled batch. 'memory': divide by the maximum weight in the memory. False: do not normalize

```
class chainerrl.replay_buffer.PrioritizedEpisodicReplayBuffer (capacity=None, alpha=0.6, beta0=0.4, betasteps=200000.0, eps=1e-08, normalize_by_max=True, default_fault_priority_func=None, uniform_ratio=0, wait_priority_after_sampling=True, return_sample_weights=True, error_min=None, error_max=None)
```



## CHAPTER 3

---

### Indices and tables

---

- `genindex`
- `modindex`
- `search`



## Symbols

`__call__()` (*chainerrl.policy.Policy* method), 31

`__call__()` (*chainerrl.q\_function.StateActionQFunction* method), 36

`__call__()` (*chainerrl.q\_function.StateQFunction* method), 36

## A

A2C (*class in chainerrl.agents*), 7

A3C (*class in chainerrl.agents*), 8

ACER (*class in chainerrl.agents*), 8

`act()` (*chainerrl.agent.Agent* method), 6

`act_and_train()` (*chainerrl.agent.Agent* method), 6

ActionValue (*class in chainerrl.action\_value*), 5

AdditiveGaussian (*class in chainerrl.explorers*), 27

AdditiveOU (*class in chainerrl.explorers*), 27

Agent (*class in chainerrl.agent*), 6

AL (*class in chainerrl.agents*), 9

`append()` (*chainerrl.replay\_buffer.ReplayBuffer* method), 43

## B

Boltzmann (*class in chainerrl.explorers*), 28

Branched (*class in chainerrl.links*), 30

## C

CategoricalDoubleDQN (*class in chainerrl.agents*), 9

CategoricalDQN (*class in chainerrl.agents*), 10

`collect_demonstrations()` (*in module chainerrl.experiments*), 23

`concatenate_recurrent_states()` (*chainerrl.links.StatelessRecurrent* method), 29

ConstantEpsilonGreedy (*class in chainerrl.explorers*), 28

ContinuousDeterministicDistribution (*class in chainerrl.distribution*), 23

ContinuousDeterministicPolicy (*class in chainerrl.policies*), 32

`copy()` (*chainerrl.distribution.Distribution* method), 21

## D

DDPG (*class in chainerrl.agents*), 10

DiscreteActionValue (*class in chainerrl.action\_value*), 5

Distribution (*class in chainerrl.distribution*), 21

DistributionalDuelingDQN (*class in chainerrl.q\_functions*), 37

DistributionalFCStateQFunctionWithDiscreteAction (*class in chainerrl.q\_functions*), 37

DistributionalSingleModelStateQFunctionWithDiscreteAction (*class in chainerrl.q\_functions*), 37

DoubleDQN (*class in chainerrl.agents*), 11

DoublePAL (*class in chainerrl.agents*), 11

DPP (*class in chainerrl.agents*), 11

DQN (*class in chainerrl.agents*), 11

DuelingDQN (*class in chainerrl.q\_functions*), 37

## E

EmpiricalNormalization (*class in chainerrl.links*), 30

entropy (*chainerrl.distribution.Distribution* attribute), 22

EpisodicReplayBuffer (*class in chainerrl.replay\_buffer*), 43

`evaluate_actions()` (*chainerrl.action\_value.ActionValue* method), 5

Explorer (*class in chainerrl.explorer*), 27

## F

FactorizedNoisyLinear (*class in chainerrl.links*), 30

FCBNDeterministicPolicy (*class in chainerrl.policies*), 32

FCBNLateActionSAQFunction (*class in chainerrl.q\_functions*), 40

FCBNQuadraticStateQFunction (*class in chainerrl.q\_functions*), 39

- FCBNSAQFunction (class in *chainerrl.q\_functions*), 40
- FCDeterministicPolicy (class in *chainerrl.policies*), 32
- FCGaussianPolicy (class in *chainerrl.policies*), 33
- FCGaussianPolicyWithFixedCovariance (class in *chainerrl.policies*), 35
- FCGaussianPolicyWithStateIndependentCovariance (class in *chainerrl.policies*), 34
- FCLateActionSAQFunction (class in *chainerrl.q\_functions*), 41
- FCLSTMDeterministicPolicy (class in *chainerrl.policies*), 33
- FCLSTMSAQFunction (class in *chainerrl.q\_functions*), 39
- FCLSTMStateQFunction (class in *chainerrl.q\_functions*), 38
- FCQuadraticStateQFunction (class in *chainerrl.q\_functions*), 38
- FCSAQFunction (class in *chainerrl.q\_functions*), 39
- FCSOftmaxPolicy (class in *chainerrl.policies*), 36
- FCStateQFunctionWithDiscreteAction (class in *chainerrl.q\_functions*), 37
- G**
- GaussianDistribution (class in *chainerrl.distribution*), 22
- GaussianHeadWithStateIndependentCovariance (class in *chainerrl.policies*), 35
- get\_recurrent\_state\_at() (chainerrl.links.StatelessRecurrent method), 29
- get\_state() (chainerrl.recurrent.Recurrent method), 41
- get\_statistics() (chainerrl.agent.Agent method), 6
- Greedy (class in *chainerrl.explorers*), 28
- greedy\_actions (chainerrl.action\_value.ActionValue attribute), 5
- I**
- IQN (class in *chainerrl.agents*), 12
- K**
- k1 (chainerrl.distribution.Distribution attribute), 22
- L**
- LinearDecayEpsilonGreedy (class in *chainerrl.explorers*), 28
- LinearInterpolationHook (class in *chainerrl.experiments*), 26
- load() (chainerrl.agent.Agent method), 6
- load() (chainerrl.replay\_buffer.ReplayBuffer method), 43
- log\_prob() (chainerrl.distribution.Distribution method), 22
- M**
- mask\_recurrent\_state\_at() (chainerrl.links.StatelessRecurrent method), 29
- max (chainerrl.action\_value.ActionValue attribute), 5
- MaxDistribution (class in *chainerrl.distribution*), 22
- MellowmaxPolicy (class in *chainerrl.policies*), 36
- MLP (class in *chainerrl.links*), 30
- MLPBN (class in *chainerrl.links*), 30
- most\_probable (chainerrl.distribution.Distribution attribute), 22
- N**
- n\_step\_forward() (chainerrl.links.StatelessRecurrent method), 29
- NatureDQNHead (class in *chainerrl.links*), 31
- NIPSDQNHead (class in *chainerrl.links*), 30
- NSQ (class in *chainerrl.agents*), 13
- P**
- PAL (class in *chainerrl.agents*), 13
- params (chainerrl.action\_value.ActionValue attribute), 5
- params (chainerrl.distribution.Distribution attribute), 22
- PCL (class in *chainerrl.agents*), 13
- PGT (class in *chainerrl.agents*), 15
- Policy (class in *chainerrl.policy*), 31
- PPO (class in *chainerrl.agents*), 16
- PrioritizedEpisodicReplayBuffer (class in *chainerrl.replay\_buffer*), 44
- PrioritizedReplayBuffer (class in *chainerrl.replay\_buffer*), 44
- prob() (chainerrl.distribution.Distribution method), 22
- Q**
- QuadraticActionValue (class in *chainerrl.action\_value*), 5
- R**
- Recurrent (class in *chainerrl.recurrent*), 41
- REINFORCE (class in *chainerrl.agents*), 17
- ReplayBuffer (class in *chainerrl.replay\_buffer*), 43
- reset\_state() (chainerrl.recurrent.Recurrent method), 41
- ResidualDQN (class in *chainerrl.agents*), 17
- S**
- sample() (chainerrl.distribution.Distribution method), 22

`sample()` (*chainerrl.replay\_buffer.ReplayBuffer method*), 43  
`sample_with_log_prob()` (*chainerrl.distribution.Distribution method*), 22  
SARSA (*class in chainerrl.agents*), 18  
`save()` (*chainerrl.agent.Agent method*), 6  
`save()` (*chainerrl.replay\_buffer.ReplayBuffer method*), 43  
`select_action()` (*chainerrl.explorer.Explorer method*), 27  
Sequence (*class in chainerrl.links*), 31  
`set_state()` (*chainerrl.recurrent.Recurrent method*), 41  
SingleActionValue (*class in chainerrl.action\_value*), 6  
SingleModelStateActionQFunction (*class in chainerrl.q\_functions*), 39  
SingleModelStateQFunctionWithDiscreteAction (*class in chainerrl.q\_functions*), 37  
SoftActorCritic (*class in chainerrl.agents*), 18  
SoftmaxDistribution (*class in chainerrl.distribution*), 22  
SoftmaxPolicy (*class in chainerrl.policies*), 36  
`state_kept()` (*in module chainerrl.recurrent*), 42  
`state_reset()` (*in module chainerrl.recurrent*), 42  
StateActionQFunction (*class in chainerrl.q\_function*), 36  
StatelessRecurrent (*class in chainerrl.links*), 28  
StatelessRecurrentBranched (*class in chainerrl.links*), 31  
StatelessRecurrentChainList (*class in chainerrl.links*), 31  
StatelessRecurrentSequential (*class in chainerrl.links*), 31  
StateQFunction (*class in chainerrl.q\_function*), 36  
StepHook (*class in chainerrl.experiments*), 26  
`stop_episode()` (*chainerrl.agent.Agent method*), 7  
`stop_episode_and_train()` (*chainerrl.agent.Agent method*), 7

## T

TD3 (*class in chainerrl.agents*), 19  
`to_factorized_noisy()` (*in module chainerrl.links*), 31  
`train_agent_async()` (*in module chainerrl.experiments*), 23  
`train_agent_batch()` (*in module chainerrl.experiments*), 24  
`train_agent_batch_with_evaluation()` (*in module chainerrl.experiments*), 24  
`train_agent_with_evaluation()` (*in module chainerrl.experiments*), 25  
TRPO (*class in chainerrl.agents*), 20