ChainerRL Documentation

Release 0.5.0

Preferred Networks, Inc.

Jan 16, 2019

Contents

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.

CHAPTER 1

Installation

1.1 How to install ChainerRL

ChainerRL is tested with Python 2.7+ and 3.5.1+. For other requirements, see `requirements.txt`.

Listing 1: requirements.txt

```
cached-property
chainer>=3.1.0
fastcache; python_version<'3.2'
funcsigs; python_version<'3.5'
future
gym>=0.9.7
numpy>=1.10.4
pillow
scipy
statistics; python_version<'3.4'
```

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
```

CHAPTER 2

API Reference

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 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 <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*) – mininum 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.**A3C**(*model*, *optimizer*, *t_max*, *gamma*, *beta=0.01*, *process_idx=0*, *phi=<function <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 regularizaiton 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 <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 regularizaiton 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.**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 <lambda>>*, *target_update_method=u'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=<logging.Logger object>*, *batch_states=<function batch_states>*)

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*

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 <lambda>>*, *target_update_method=u'hard'*, *soft_update_tau=0.01*, *n_times_update=1*, *average_q_decay=0.999*, *average_loss_decay=0.99*, *batch_accumulator=u'mean'*, *episodic_update=False*, *episodic_update_len=None*, *logger=<logging.Logger object>*, *batch_states=<function batch_states>*)

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*, *update_interval=1*, *target_update_interval=10000*, *clip_delta=True*, *phi=<function <lambda>>*, *target_update_method=u'hard'*, *soft_update_tau=0.01*, *n_times_update=1*, *average_q_decay=0.999*, *average_loss_decay=0.99*, *batch_accumulator=u'mean'*, *episodic_update=False*, *episodic_update_len=None*, *logger=<logging.Logger object>*, *batch_states=<function batch_states>*)

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** (*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*

class chainerrl.agents.**NSQ**(*q_function*, *optimizer*, *t_max*, *gamma*, *i_target*, *explorer*, *phi=<function <lambda>>*, *average_q_decay=0.999*, *logger=<logging.Logger object>*, *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 <lambda>>*, *pi_loss_coef=1.0*, *v_loss_coef=0.5*, *rollout_len=10*, *batchsize=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*, *average_entropy_decay=0.999*, *average_value_decay=0.999*, *explorer=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 regularizaiton 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 <lambda>>*, *target_update_method=u'hard'*, *soft_update_tau=0.01*, *n_times_update=1*, *average_q_decay=0.999*, *average_loss_decay=0.99*, *logger=<logging.Logger object>*, *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.**REINFORCE**(*model*, *optimizer*, *beta=0*, *phi=<function <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 regularizaiton 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 <lambda>>*, *target_update_method=u'hard'*, *soft_update_tau=0.01*, *n_times_update=1*, *average_q_decay=0.999*, *average_loss_decay=0.99*, *batch_accumulator=u'mean'*, *episodic_update=False*, *episodic_update_len=None*, *logger=<logging.Logger object>*, *batch_states=<function batch_states>*)

SARSA.

Unlike DQN, this agent uses actions that have been actually taken to compute target Q values, thus is an on-policy algorithm.

2.3 Distributions

2.3.1 Distribution interfaces

class chainerrl.distribution.**Distribution**
Batch of distributions of data.

copy(*x*)
Copy a distribion 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

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*)
Continous deterministic distribution.

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

2.4 Experiments

2.4.1 Training and evaluation

chainerrl.experiments.**train_agent_async**(*outdir*, *processes*, *make_env*, *profile=False*, *steps=80000000*, *eval_interval=1000000*, *eval_n_runs=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_runs** (*int*) – Number of runs for each time of evaluation.
- **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** (*list*) – List 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_with_evaluation**(*agent*, *env*, *steps*, *eval_n_steps*, *eval_n_episodes*, *eval_interval*, *outdir*, *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.
- **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** (*list*) – List 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.

2.4.2 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 Using recurrent models

2.5.1 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.

2.5.2 Utilities

chainerrl.recurrent.**state_kept**(**args*, ***kwds*)

Keeps the previous state of a given link.

This is a context manager that saves 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**(**args*, ***kwds*)

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

This is a context manager that saves 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
```

CHAPTER 3

Indices and tables

- genindex
- modindex
- search

Index

T

U