Noname manuscript No.
(will be inserted by the editor)

Extensions for DDPG and analysis of its components

subtitle here

Maximilian Gehrke · Tabea Wilke · Yannik Frisch

Received: date / Accepted: date

Abstract TODO

Keywords DDPG · DQN · DPG

1 Introduction

Deep Deterministic Policy Gradients (DDPG) arises from Deterministic Policy Gradients (DPG) and Deep Q-Learning (DQN). In the following we describe the underlying algorithms DPG and DQN and which aspects DDPG uses of both of them.

1.1 Deterministic Policy Gradient (DPG)

1.2 Deep Q-Learning (DQN)

DQN is the combination of neural networks and q-learning. It works on a deterministic environment with the goal to achieve the optimal action-value function. This means finding the best action with respect to the rewards also in the future to a given state. In terms of a formula it is represented by

$$Q^*(s_t, a_t) = \pi$$

with λ as discount factor smaller but close to 1, so the agent takes also future reward into account. The rewards of the future will have impact on the result

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

Table 1 Please write your table caption here

first	second	third
number	number	number
number	number	number

but the influence decreases with time. For estimating the action-value function a deep network is used. Furthermore, a replay buffer is used which will save samples of the environment. Therefore, it is possible to achieve a non correlated batch.

- DQN uses deep networks to estimate the action-value function
  - it can only handle discrete and low-dim action spaces
- discretizing the action space often suffers from the course of dimensionality
- PolicyGradientTheorem from continous space to discrete space presented in DPG paper
- naive extension of DPG with nns turns out to be unstable for challenging problems
- Deep DPG (DDPG): combination of DQN and DPG, where:
  - networks are trained off-policy with samples from a replay buffer to minimize the temporal correlations between samples
  - the networks are trained with target networks to give consistent targets during temporal difference backups
  - batch normalization is used
- DDPG is able to learn from low dim observations (torques etc.), aswell as from high dim observations in pixel space

Your text comes here. Separate text sections with

2 Extensions to the Algorithm

We propose several possible extensions and show their performance on a task. Text with citations [2] and [1].

2.1 Subsection title

as required. Don't forget to give each section and subsection a unique label (see Sect. 2).

Paragraph headings Use paragraph headings as needed.

$$a^2 + b^2 = c^2 \tag{1}$$

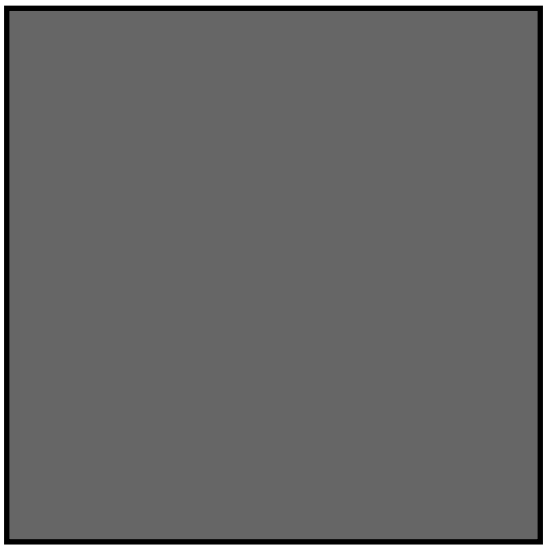

Fig. 1 Please write your figure caption here

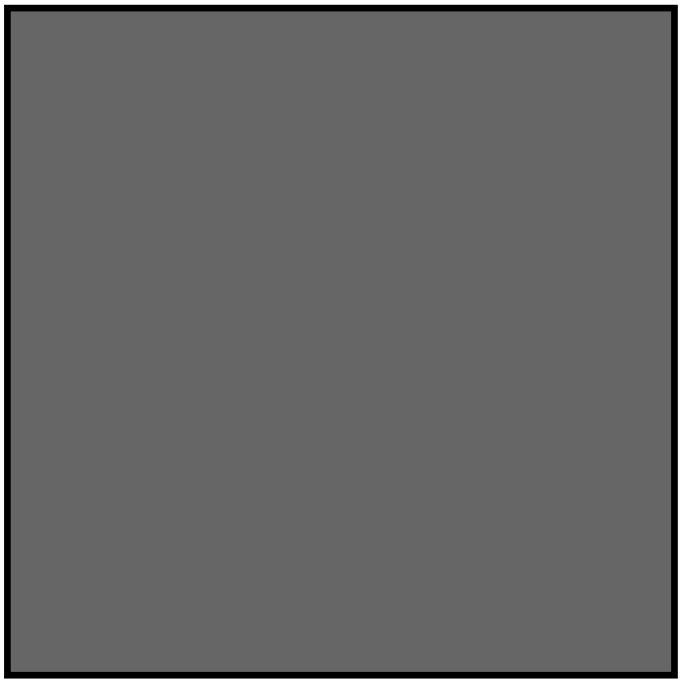

Fig. 2 Please write your figure caption here

References

1. Author, Article title, Journal, Volume, page numbers (year)
2. Author, Book title, page numbers. Publisher, place (year)