RL seminar #5: DDPG & Bandits

...

MIPT, Deep learning lab & iPavlov.ai

2 Dec 2017

Outline

- Class information
    - Assignments
- Deterministic Policy Gradient
    - Overview
    - Notation
    - Algorithm
    - alg
- DDPG ref
    - ddgg refs
- Multi-armed bandits
    - Overview
    - Upper confidence bound
    - Applications
- Questions
- Next steps

Table of Contents

- Class information
  - Assignments
- Deterministic Policy Gradient
  - Overview
  - Notation
  - Algorithm
  - alg
- DDPG ref
  - ddgg refs
- Multi-armed bandits
  - Overview
  - Upper confidence bound
  - Applications
- Questions
- Next steps

Assignments

Coding (Programming assignment #2)

Deadline: ... Dec 2017

Quiz

Quiz N4 to be issued.

Questions

...

Course

...

Table of Contents

- Class information
  - Assignments
- Deterministic Policy Gradient
  - Overview
  - Notation
  - Algorithm
  - alg
- DDPG ref
  - ddgg refs
- Multi-armed bandits
  - Overview
  - Upper confidence bound
  - Applications
- Questions
- Next steps

DPG: Overview

Task

Continuous actions in high dimensions

Motivation

- Deterministic policy gradient can be estimated much more efficiently than the usual stochastic policy gradient.
- Deterministic policy gradient is simple and natural

from David Silver's lecture on Deterministic Policy Gradient Algorithms

Nontation

- Continuous state space $s \in \mathbb{R}^l$
- Continuous action space $a \in A^m$
- Parameter vector $\theta \in \mathbb{R}^n$
- Deterministic policy $a = \mu_\theta(s)$
- Stochastic policy $\pi_\theta(s, a) = p(a|s; \theta)$

Find parameters θ optimizing performance of policy

from David Silver's lecture on Deterministic Policy Gradient Algorithms

Policy Gradient theorem

Stochastic case:

$$J(\theta) = \mathbb{E}_{s \sim p^{\pi}(s)} \Big[\int_a \pi_\theta(s,a) R(s,a) da \Big]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{s,a} \Big[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi}(s,a) \Big]$$

Deterministic case:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s} \Big[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu}(s,a)|_{a=\mu_\theta}(s) \Big]$$

Update policy in the direction that most improves Q

Algorithm 1 DDPG algorithm

```
Randomly initialize critic n-k $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights $\theta^Q$ and $\theta^\mu$.
Initialize target network $Q'$ and $\mu'$ with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer $R$
for episode = 1, M do
    Initialize a random process $\mathcal{N}$ for action exploration
    Receive initial observation state $s_1$
    for t = 1, T do
        Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ by current policy and exploration noise
        Execute action $a_t$ and observe reward $r_t$ and observe new state $s_{t+1}$
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $R$
        Sample a random minibatch of $N$ transitions $(s_i, a_i, r_i, s_{i+1})$ from $R$
        Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
        Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
        Update the actor policy using the sampled policy gradient:

            $\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$

        Update the target networks:

            $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$

    end for
end for
```

Table of Contents

Class information
- Assignments

Deterministic Policy Gradient
- Overview
- Notation
- Algorithm
- alg

DDPG ref
- ddgg refs

Multi-armed bandits
- Overview
- Upper confidence bound
- Applications

Questions

Next steps

DDPG code

- https://github.com/fgvbrt/nips˙rl
- https://github.com/Scitator/Run-Skeleton-Run
- https://github.com/vy007vikas/PyTorch-ActorCriticRL

Table of Contents

- Class information
  - Assignments
- Deterministic Policy Gradient
  - Overview
  - Notation
  - Algorithm
  - alg
- DDPG ref
  - ddgg refs
- Multi-armed bandits
  - Overview
  - Upper confidence bound
  - Applications
- Questions
- Next steps

MA bandits: overview

Task definition

…

Upper confidence bound

Optimal exploration

...

Applications

…

Table of Contents

- Class information
  - Assignments
- Deterministic Policy Gradient
  - Overview
  - Notation
  - Algorithm
  - alg
- DDPG ref
  - ddgg refs
- Multi-armed bandits
  - Overview
  - Upper confidence bound
  - Applications
- Questions
- Next steps

Questions

1. L12 – 'Closer look at backward pass'.
2. L12 – 'Benefits of soft optimality'.
3. L12 – 'More efficient sample-based updates / Importance sampling'.

Table of Contents

- Class information
  - Assignments
- Deterministic Policy Gradient
  - Overview
  - Notation
  - Algorithm
  - alg
- DDPG ref
  - ddgg refs
- Multi-armed bandits
  - Overview
  - Upper confidence bound
  - Applications
- Questions
- Next steps

Next steps

Plan for the week

- Quiz N4 (?)
- Home assignment N3 (?)

Reading

Lectures 15-17 of CS294.
Please, post your questions about lectures in google doc:
`https://goo.gl/qN6jmJ`

Rating

ref: `goo.gl/yxqhBg`
HW scores are coming (week 4-10 Dec)