Machine Learning for Robotics
Intelligent Systems Series
Lecture 12

Georg Martius

MPI for Intelligent Systems, Tübingen, Germany

July 17, 2017

Reminder of Actor critics and some tricks what people use.

- Example A3C (Asynchronous Advantage Actor Critic)
  `https://arxiv.org/abs/1602.01783`
- See "Deep RL Tutorial NIPS 2016" by D. Silver Slides 37ff.

To compute derivatives of networks:

- see Lecture 5. Slides 24 onwards
- or use **Automatic Differentiation**
- e.g. implemented in tensorflow, theano and many more.
- these frameworks also implement improved stochastic gradient methods, such as RMS-Prop, Adam etc.
- Tensorflow: very short tutorial:
  `http://cv-tricks.com/artificial-intelligence/deep-learning/deep-learning-frameworks/tensorflow-tutorial`

PGPE: Policy gradient by parameter exploration

by Sehnke, Osendorfer , Rückstieß, Graves, Peters and Schmidhuber, 2010

- Idea: instead of exploring the actions, explore the parameters of the policy
- Go through the math in the paper on the blackboard
  (paper can be downloaded from course website)
- Here are some slides:
  `http://boemund.dagstuhl.de/mat/Files/11/11131/11131.SehnkeFrank.Slides1.pdf`

- Key features: each episode the parameters of a deterministic policy are sampled from a (normal) distribution.
- The parameters of that distribution are updated according to gained rewards
- Policy can be non-differentiable
- Drawback: needs also many episodes (but might explore better than action-exploring PG methods)

PGPE – data reuse

- Version with data reuse: "Efficient Sample Reuse in Policy Gradients with Parameter-based Exploration" by Zhao et al. 2013, `https://arxiv.org/abs/1301.3966`
- Idea: weight previously collected data according to importance sampling
- introduce new baseline (coping with this importance sampling)