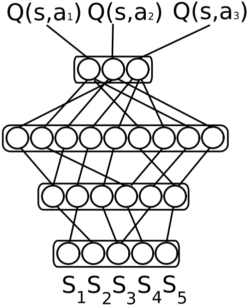
From Policy Gradient to Actor-Critic methods

8. Deep Deterministic Policy Gradient (and TD3)

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

The Q-network in DQN



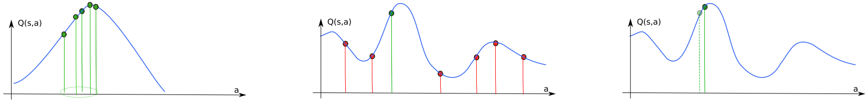
- Parametrized representation of the critic $\hat{Q}^{\pi_\theta}_{\phi}(\mathbf{s}_t, \mathbf{a}_t)$
- Q-network equivalent to the Q-Table (with an infinity of state rows)
- For each observed $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$:

$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow Q(\mathbf{s}_t, \mathbf{a}_t) + \alpha[r_t + \gamma \max_{\mathbf{a} \in A} Q(\mathbf{s}_{t+1}, \mathbf{a}) - Q(\mathbf{s}_t, \mathbf{a}_t)]$$

- Select action by finding $\max_{\mathbf{a} \in A} Q(\mathbf{s}, \mathbf{a})$ (as in Q-LEARNING)
- Limitation: requires one output neuron per action

Moving to continuous actions

- Two things become too hard:
  - Selecting actions by finding $\max_{\mathbf{a} \in A} Q(\mathbf{s}, \mathbf{a})$
  - Computing $\max_{\mathbf{a} \in A} Q(\mathbf{s}_{t+1}, \mathbf{a})$ in the update rule

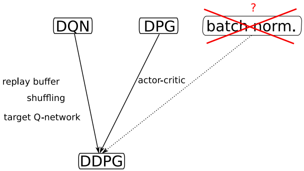


- Three classes of solutions
  1. Use an easily optimized model (e.g. convex) (NAF, Wang et al. 2016)
  2. Sample a limited set of actions (QT-Opt, Kalashnikov et al., 2018)
  3. DDPG: train a side estimator of the best action (also true of SAC)

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016

DDPG: ancestors

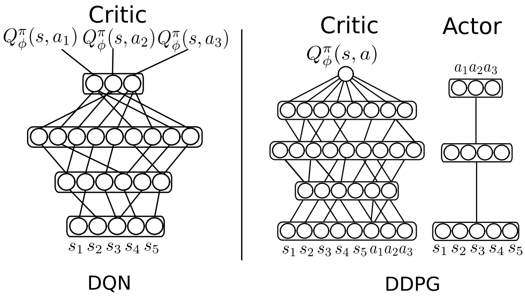


- Most of the actor-critic theory for continuous problem is for stochastic policies (policy gradient theorem, compatible features, etc.)
- DPG: an efficient gradient computation for deterministic policies, with proof of convergence
- Batch norm: inconclusive studies about importance
- Used on 32 classic control benchmarks, sometimes from pixels

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014) Deterministic policy gradient algorithms. In *ICML*

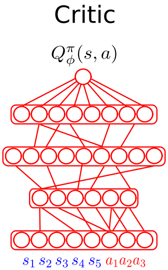
Ioffe, S. & Szegedy, C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*

General architecture



- Actor $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$, critic $\hat{Q}^{\pi_\theta}_\phi(\mathbf{s}_t, \mathbf{a}_t)$ (a single output neuron)
- All updates based on SGD
- Adaptive gradient descent techniques tune the step size (RProp, RMSProp, Adagrad, Adam...)

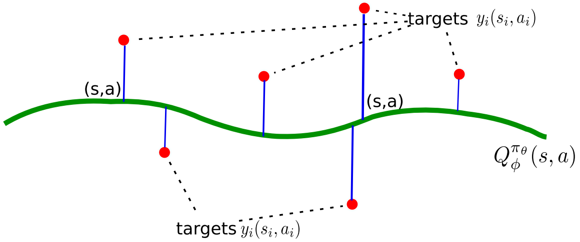
Training the critic



- Same idea as in DQN, but for actor-critic rather than Q-LEARNING
- Supervised learning: minimize $L(\phi) = (y^*(\mathbf{s},\mathbf{a}) - \hat{F}_{\phi}(\mathbf{s}_i,\mathbf{a}_i|\phi))^2$
- For each sample i, the Q-network should minimize the RPE: $\delta_t = r_t + \gamma \hat{Q}^{\pi\theta}_{\phi}(s_{t+1}, \pi(s_{t+1})) - \hat{Q}^{\pi\theta}_{\phi}(\mathbf{s}_t,\mathbf{a}_t)$
- Given a minibatch of N samples $\{\mathbf{s}_i,\mathbf{a}_i,r_i,\mathbf{s}_{i+1}\}$ and a target network Q', compute $y_i = r_i + \gamma \hat{Q}'^{\pi\theta}_{\phi'}(\mathbf{s}_{i+1}, \pi(\mathbf{s}_{i+1}))$
- And update θ by minimizing the loss function

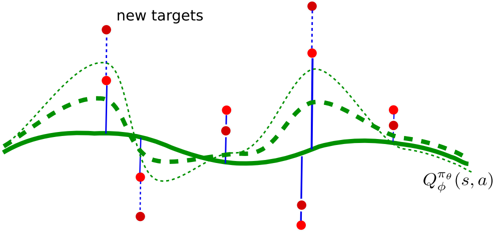
$$L = 1/N \sum_i (y_i - \hat{Q}^{\pi\theta}_{\phi}(\mathbf{s}_i,\mathbf{a}_i|\phi))^2 \quad (1)$$

Learning the neural Q-function



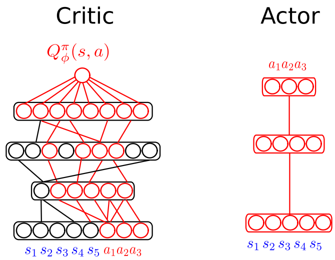
- In the tabular case, each Q-value is updated separately
- In the continuous function approximation setting, interdependencies
- Thus update θ by minimizing the (squared error) loss function

Trick 1: Stable Target Q-function



- The target $y_i = r_i + \gamma \max_a \hat{Q}_\phi^{\pi_\theta}(\mathbf{s}_{i+1}, a)|\phi)$ is itself a function of Q
- Thus this is not truly supervised learning, and this is unstable
- Key idea: "periods of supervised learning"
- Compute the loss function from a separate *target critic* $\hat{Q}'^{\pi_\theta}_{\phi'}(...|\phi')$
- So rather compute $y_i = r_i + \gamma \max_a \hat{Q}'^{\pi_\theta}_{\phi'}(\mathbf{s}_{i+1}, a|\phi')$
- In DQN, ϕ' is updated to ϕ only each K iterations
- In DDPG, update ϕ' using $\phi' \leftarrow (1-\tau)\phi' + \tau\phi$ with a small gain τ

Training the actor



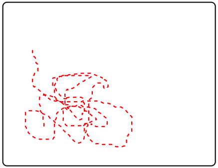
- Deterministic policy gradient theorem: the true policy gradient is

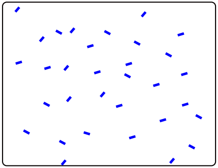
$$\nabla_{\theta}\pi(\mathbf{s}_t,\mathbf{a}_t) = \mathbb{E}_{\mathbf{s}_t,\mathbf{a}_t\sim\pi_{\theta}(.)}[\nabla_a \hat{Q}^{\pi_\theta}_{\phi}(\mathbf{s}_t,\mathbf{a}_t)\nabla_{\theta}\pi(s|\theta)] \quad (2)$$

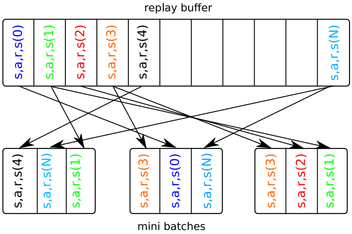
- $\nabla_a \hat{Q}^{\pi_\theta}_{\phi}(\mathbf{s}_t,\mathbf{a}_t)$ is used as error signal to update the actor weights.
- Comes from NFQCA
- $\nabla_a \hat{Q}^{\pi_\theta}_{\phi}(\mathbf{s}_t,\mathbf{a}_t)$ is a gradient over actions
- $y = f(w.x+b)$ (symmetric roles of weights and inputs)
- Gradient over actions $\sim$ gradient over weights

Hafner, R. & Riedmiller, M. (2011) Reinforcement learning in feedback control. *Machine learning*, 84(1-2), 137–169.

Trick2: Replay buffer shuffling


Non i.i.d. samples

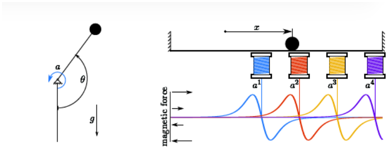

i.i.d. samples



- Agent samples are not independent and identically distributed (i.i.d.)
- Shuffling a replay buffer (RB) makes them more i.i.d.
- It improves a lot the sample efficiency
- Recent data in the RB come from policies close to the current one

Lin, L.-J. (1992) Self-Improving Reactive Agents based on Reinforcement Learning, Planning and Teaching. *Machine Learning*, 8(3/4), 293–321

Replay buffer management



(a) Pendulum task (b) Magnetic manipulation task

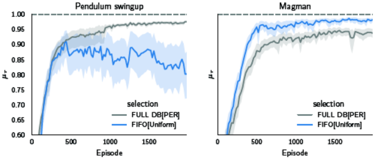
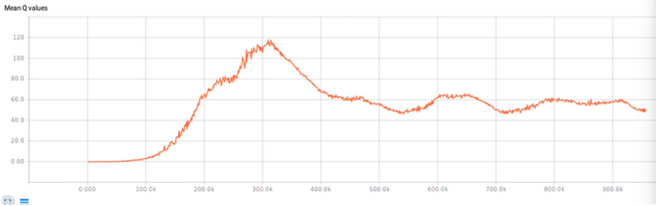


Figure 3: Comparison of the state-of-the-art (FULL DB[PER]) and the default method (FIFO[Uniform]) for experience selection on our two benchmark problems.

- Different replay buffer management strategies are optimal in different problems

de Bruin, T., Kober, J., Tuyls, K., & Babuška, R. (2018) Experience selection in deep reinforcement learning for control. *Journal of Machine Learning Research*, 19(9):1–56

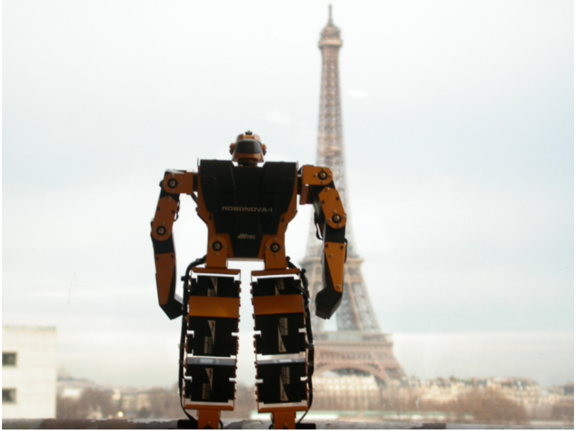
Twin Delayed DPG



- All descendants of Q-learning suffer from over-estimation bias
- Clipping the target critic from the knowledge of R_{max} helps
- TD3: Introduce two critics $\hat{Q}^{\pi_\theta}_{\phi_1}$ and $\hat{Q}^{\pi_\theta}_{\phi_2}$
- Compute the TD-target as the minimum to reduce the over-estimation bias
- Less problem knowledge than target critic clipping
- Next lesson: Soft Actor Critic

Fujimoto, S., van Hoof, H., & Meger, D. (2018) Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*

Any question?



Send mail to: `Olivier.Sigaud@upmc.fr`

Tim de Bruin, Jens Kober, Karl Tuyls, and Robert Babuška.
Experience selection in deep reinforcement learning for control.
Journal of Machine Learning Research, 19(9):1–56, 2018.

Scott Fujimoto, Herke van Hoof, and Dave Meger.
Addressing function approximation error in actor-critic methods.
arXiv preprint arXiv:1802.09477, 2018.

Roland Hafner and Martin Riedmiller.
Reinforcement learning in feedback control.
Machine learning, 84(1-2):137–169, 2011.

Sergey Ioffe and Christian Szegedy.
Batch normalization: Accelerating deep network training by reducing internal covariate shift.
arXiv preprint arXiv:1502.03167, 2015.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al.
Qt-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation.
arXiv preprint arXiv:1806.10293, 2018.

Long-Jin Lin.
Self-Improving Reactive Agents based on Reinforcement Learning, Planning and Teaching.
Machine Learning, 8(3/4):293–321, 1992.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller.
Deterministic policy gradient algorithms.
In *Proceedings of the 30th International Conference in Machine Learning*, 2014.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas.
Sample efficient actor-critic with experience replay.
arXiv preprint arXiv:1611.01224, 2016.