DDPG

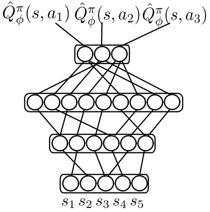
Olivier Sigaud

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

The Q-network in DQN

state / action	a_0	a_1	a_2	a_3
$\mathbf{s}_0$	0.66	0.88*	0.81	0.73
$\mathbf{s}_1$	0.73	0.63	0.9*	0.43
$\mathbf{s}_2$	0.73	0.9	0.95*	0.73
$\mathbf{s}_3$	0.81	0.9	1.0*	0.81
$\mathbf{s}_4$	0.81	1.0*	0.81	0.9
$\mathbf{s}_5$	0.9	1.0*	0.0	0.9

Critic



DQN

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

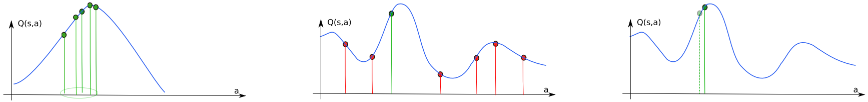
$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow Q(\mathbf{s}_t, \mathbf{a}_t) + \alpha[r_t + \gamma \max_{\mathbf{a} \in A} Q(\mathbf{s}_{t+1}, \mathbf{a}) - Q(\mathbf{s}_t, \mathbf{a}_t)]$$

- 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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.

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

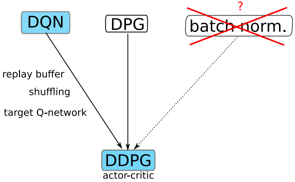


- Three classes of solutions
  1. Use an easily optimized model (e.g. convex) (NAF, Wang et al. 2016)
  2. Sample a limited set of actions (QT-Opt, Kalashnikov et al., 2018)
  3. DDPG: train a side estimator of the best action (also true of SAC)

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

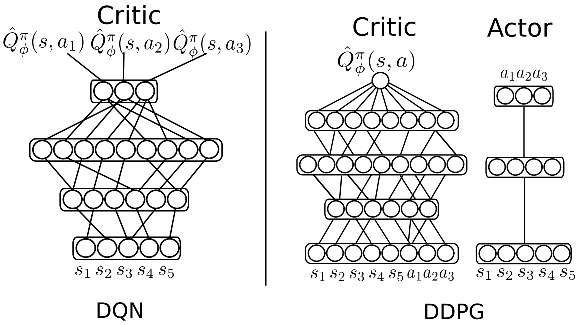


- Most of the actor-critic theory for continuous problem is for stochastic policies (policy gradient theorem, compatible features, etc.)
- DPG: an efficient gradient computation for deterministic policies, with proof of convergence
- Batch norm: inconclusive studies about impact
- Used on 32 classic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

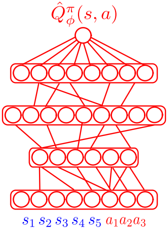


- Actor $\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$, critic $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t,\mathbf{a}_t)$ (a single output neuron)
- All updates based on SGD
- Adaptive gradient descent techniques tune the step size (RProp, RMSProp, Adagrad, Adam...)

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015) Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971 7/9/15*

Training the critic

Critic

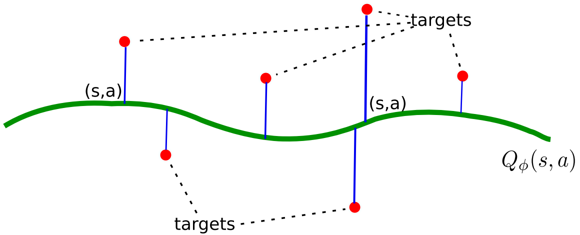


- Same idea as in DQN, but for actor-critic rather than Q-LEARNING

- Supervised learning: minimize $L(\boldsymbol{\phi}) = (y^*(\mathbf{s},\mathbf{a}) - \hat{F}_{\boldsymbol{\phi}}(\mathbf{s}_i,\mathbf{a}_i|\boldsymbol{\phi}))^2$
- For each sample i, the Q-network should minimize the RPE:
  $\delta_t = r_t + \gamma \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}, \pi_{\boldsymbol{\theta}}(\mathbf{s}_{t+1})) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t,\mathbf{a}_t)$
- Given a minibatch of N samples $\{\mathbf{s}_i,\mathbf{a}_i,r_{i+1},\mathbf{s}_{i+1}\}$ and a target network Q', compute $y_i = r_{i+1} + \gamma \hat{Q}'^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}'}(\mathbf{s}_{i+1}, \pi(\mathbf{s}_{i+1}))$
- And update ϕ by minimizing the loss function

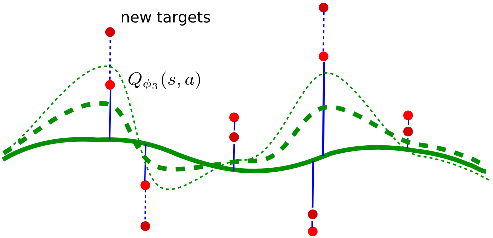
$$L = 1/N \sum_i (y_i - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_i,\mathbf{a}_i|\boldsymbol{\phi}))^2$$

Learning the neural Q-function



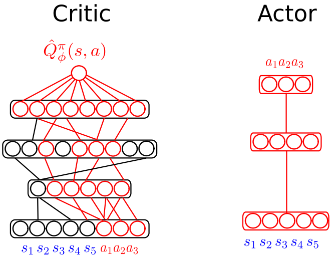
- In the tabular case, each Q-value is updated separately
- In the continuous state and action setting, interdependencies between updates
- Thus update ϕ by minimizing the squared TD loss function over minibatches

Trick 1: Stable Target Q-function



- The target $y_i = r_{i+1} + \gamma \max_a \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_{i+1}, a)|\boldsymbol{\phi})$ is itself a function of $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}$
- 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* $\hat{Q}'^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}'}(...|\boldsymbol{\phi}')$
- So rather compute $y_i = r_{i+1} + \gamma \max_a \hat{Q}'^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}'}(\mathbf{s}_{i+1}, a|\boldsymbol{\phi}')$
- In DQN, $\boldsymbol{\phi}'$ is updated to $\boldsymbol{\phi}$ only each K iterations
- In DDPG, update $\boldsymbol{\phi}'$ using $\boldsymbol{\phi}' \leftarrow (1-\tau)\boldsymbol{\phi}' + \tau\boldsymbol{\phi}$ with a small gain τ

Training the actor



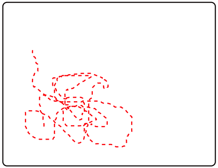
- Deterministic policy gradient theorem (Silver et al. 2014): the policy gradient is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_t,\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.)}[\nabla_a \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t,\mathbf{a}_t) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\mathbf{s}_t)] \tag{1}$$

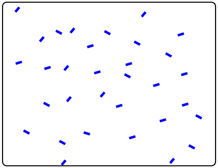
- $\nabla_a \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t,\mathbf{a}_t)$ is used as error signal to update the actor weights.
- Comes from NFQCA
- $\nabla_a \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t,\mathbf{a}_t)$ is a gradient over actions
- $y = f(w.x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

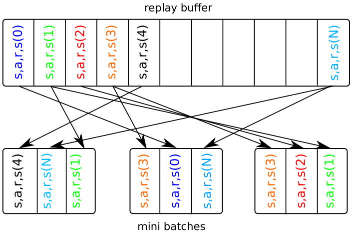
Trick2: Replay buffer shuffling



Non i.i.d. samples



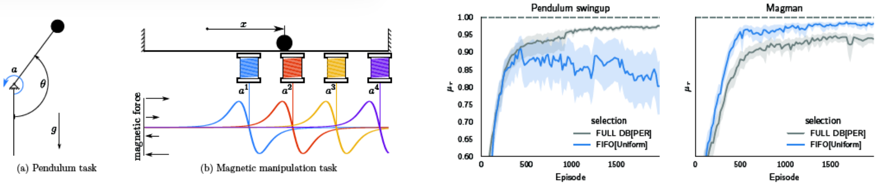
i.i.d. samples



- 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

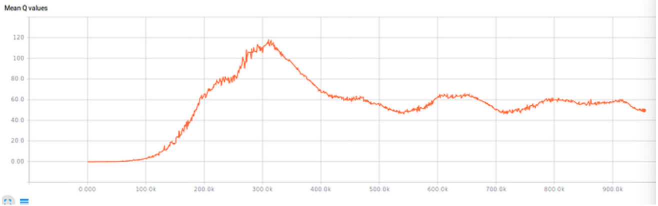
Replay buffer management



- 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

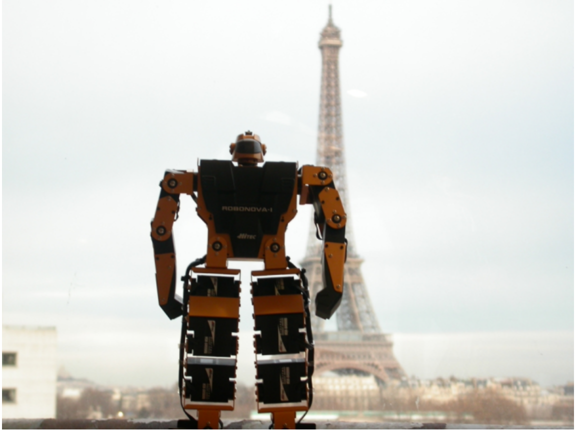
Twin Delayed Deep Deterministic PG



- All descendants of Q-learning suffer from over-estimation bias
- Clipping the critic from the knowledge of R_{max} helps
- TD3: Introduce two critics $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}_1}$ and $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}_2}$
- Compute the TD-target as the minimum to reduce the over-estimation bias
- Less problem knowledge than 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@isir.upmc.fr`

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

Fujimoto, S., van Hoof, H., and Meger, D. (2018).
Addressing function approximation error in actor-critic methods.
In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1582–1591. PMLR.

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

Ioffe, S. and Szegedy, C. (2015).
Batch normalization: Accelerating deep network training by reducing internal covariate shift.
In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. (2018).
QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation.
arXiv preprint arXiv:1806.10293.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016).
Continuous control with deep reinforcement learning.
In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015).
Human-level control through deep reinforcement learning.
Nature, 518(7540):529–533.

ISIR
INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. A. (2014).
Deterministic policy gradient algorithms.
In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 387–395. JMLR.org.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2017).
Sample efficient actor-critic with experience replay.
In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.