Modern Policy Search

An overview

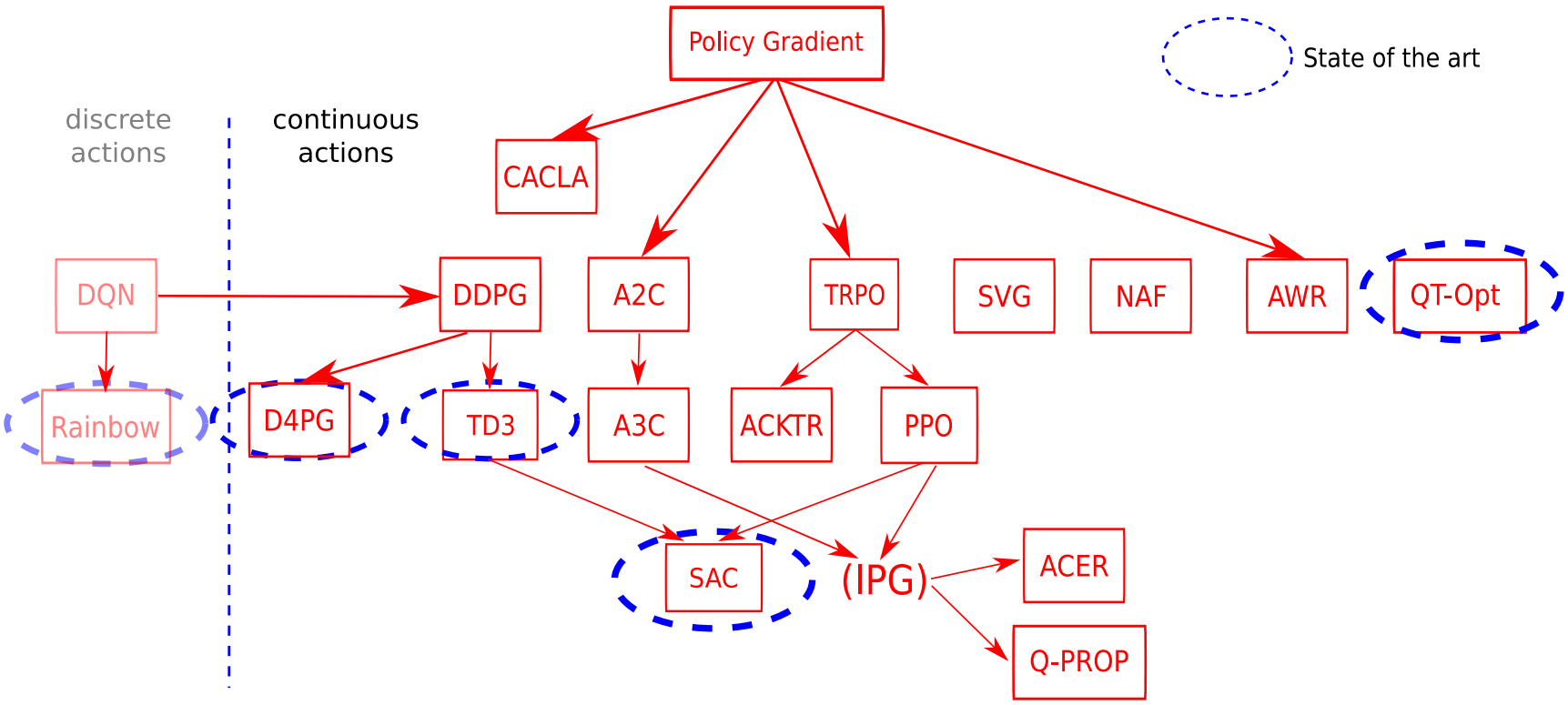
Olivier Sigaud

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

Outline

- This video comes after a policy gradient class available here: https://www.youtube.com/watch?v=_RQYWSvMyyc
- Overview of SOTA RL algorithms in the continuous action case
- Each algorithm is presented in one slide, with a link to a dedicated video
- Then different perspectives are used to relate those algorithms to each other:
  - On-policy versus Off-policy
  - The deadly triad perspective
  - The continuous action Q-learning perspective
  - Distributional RL
- A few general conclusions are given

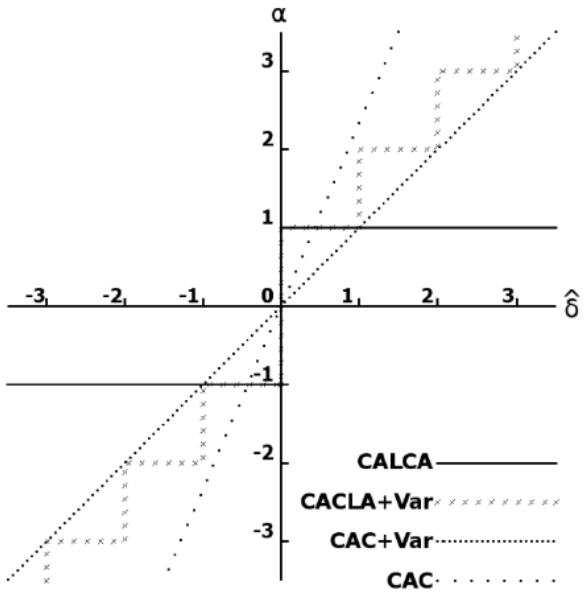
Overview



- Continuous action algorithms every RL student should know:
  A3C, DDPG, TD3, PPO, SAC

CACLA

Hado Van Hasselt and Marco A. Wiering (2007) Reinforcement learning in continuous action spaces. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 272–279.



- CACLA: an important continuous action ancestor, V as critic, stochastic actor, update only if $\delta > 0$

A2C, A3C

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu (2016) Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*

- Straightforward implementation of actor-critic policy gradient approach
- Learns the value function V as a critic through TD estimate
- Adds an entropy regularization to favor exploration
- Gradient: $\nabla_{\theta'} \log \pi(a_t|s_t;\theta')(R_t - V(s_t;\theta_v)) + \beta \nabla_{\theta'} H(\pi(s_t;\theta'))$
- A3C: Asynchronous A2C with many actors
- Does not use a replay buffer (asynchrony ensures more i.i.d. samples)
- Uses N-step return (unclear use of memory)
- Designed to be on-policy

TRPO

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015) Trust Region Policy Optimization. *CoRR, abs/1502.05477*

- MC estimate approach to policy gradient
- Uses a surrogate gradient with important sampling
- Enforces a natural gradient constraint to maintain exploration while keeping close to previous policy
- Conjugate gradient optimization approach
- See my dedicated video: `https://www.youtube.com/watch?v=qh_qTNG3UVg`

PPO, ACKTR

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347.

- Two PPO algorithms: forget the conjugate gradient approach
  - Algo 1: Soft constraint (regularization) on trust region (can use SGD)
  - Algo 2: Clipped importance sampling loss function (again, can use SGD)
  - GAE: Use N-step or λ return instead of Monte Carlo

Yuhuai Wu, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba (2017) Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. *arXiv preprint arXiv:1708.05144*

- K-FAC: Kronecker Factored Approximated Curvature: efficient estimate of natural gradient
- Uses block diagonal estimations of the Hessian matrix, to do better than first order
- ACKTR: TRPO with K-FAC natural gradient calculation
- See my dedicated video: https://www.youtube.com/watch?v=qh_qTNG3UVg

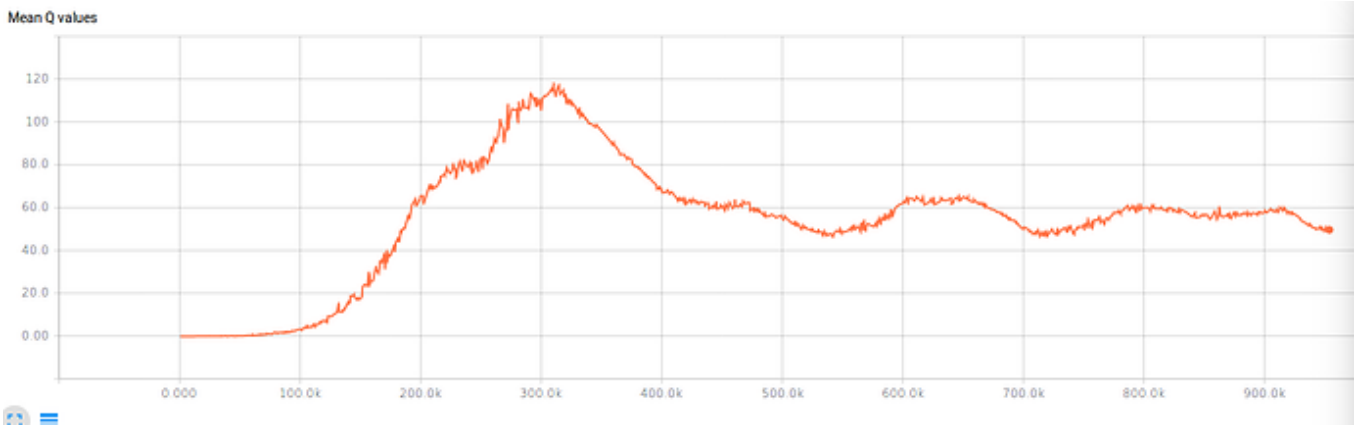
DDPG

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*

- Uses a deterministic policy, and the deterministic policy gradient
- Reuses several tricks from DQN: replay buffer shuffling, target networks
- Suffers from over-estimation bias
- Somewhat off-policy, sample efficient and unstable
- See my dedicated video:
  `https://www.youtube.com/watch?v=_pbd6TCjmaw`

TD3

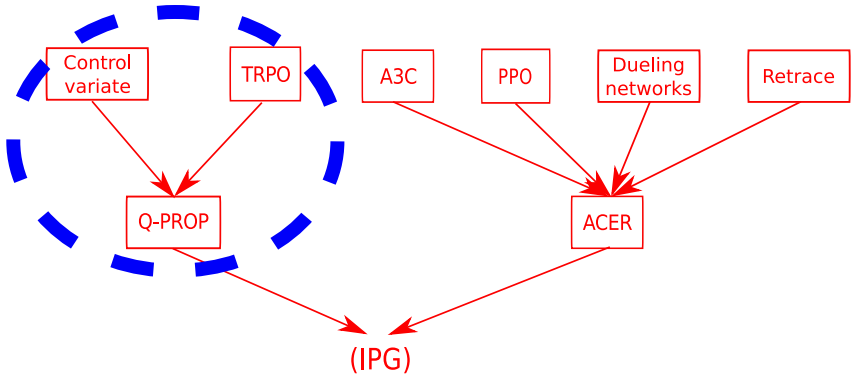
Fujimoto, S., van Hoof, H., & Meger, D. (2018) Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*



- Builds on DDPG
- Uses two critics and takes the min to counter over-estimation bias
- Performs on par with SAC (sometimes better, sometimes worse)
- Depends on the environment: impact of entropy regularization

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pp. 151–160, 2019

Q-PROP

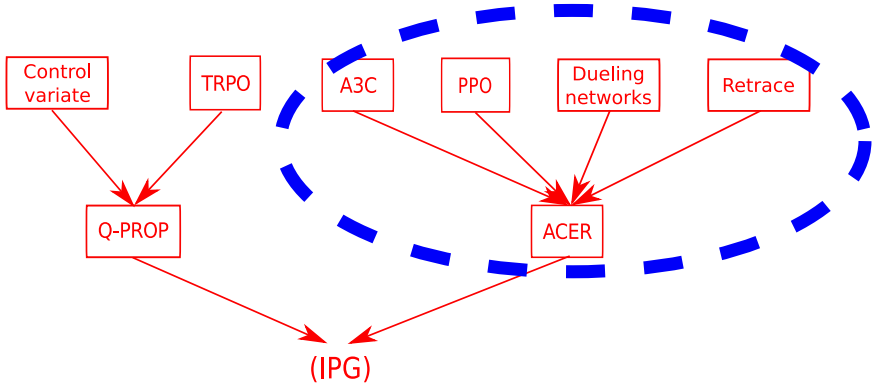


Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247*, 2016

- Integrates deterministic and stochastic policy gradient using a *control variate* formalization
- Integrates (on-policy) MC policy gradient methods and (off-policy) actor-critic methods into a single framework
- Performs one or the other depending on some hyperparameters

John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012

ACER



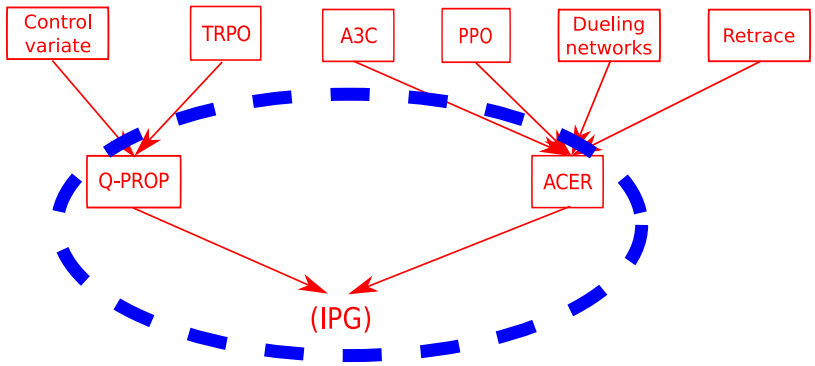
Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016

- A3C + *truncated importance sampling with bias correction* (see PPO)
- Stochastic dueling network architecture to efficiently approximate the advantage function
- Uses RETRACE to perform off-policy n-step return updates

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2016

The Interpolated Policy Gradient framework



Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, Bernhard Schölkopf, and Sergey Levine (2017)
Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning. *arXiv preprint arXiv:1706.00387*

- Merging on-policy and off-policy methods
- Too complicated, too many hyperparameters

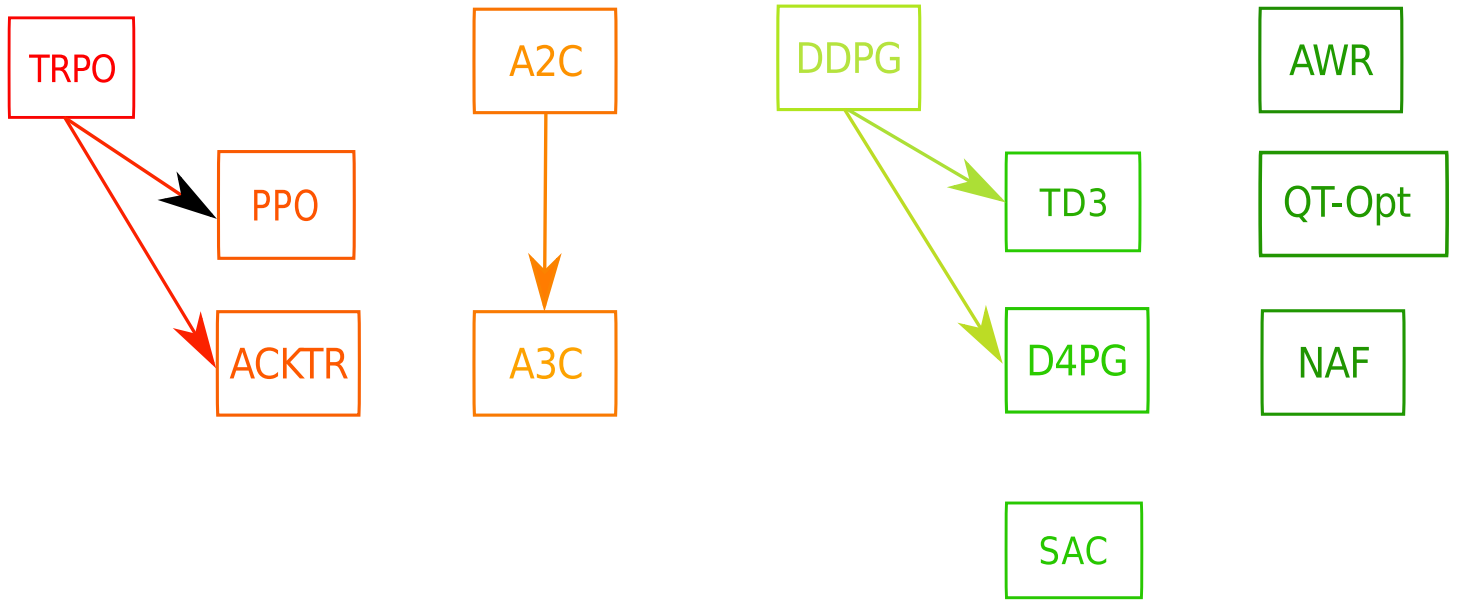
Soft Actor Critic

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. (2018) Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*

- Uses a stochastic policy
- Integrates entropy regularization inside the objective function
- In contrast to A2C, which adds entropy to the chosen action
- Uses two critics as TD3
- Uses N-step return
- See my dedicated video:
  https://www.youtube.com/watch?v=_nFX0Zpo50U

Off-policy versus On-policy



- On-policy is generally more stable, but requires more sample
- Off-policy is more sample efficient, but more unstable
- A lot of current research is dedicated to finding the best of both worlds
- SAC is doing well under this perspective

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. (2020) What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*

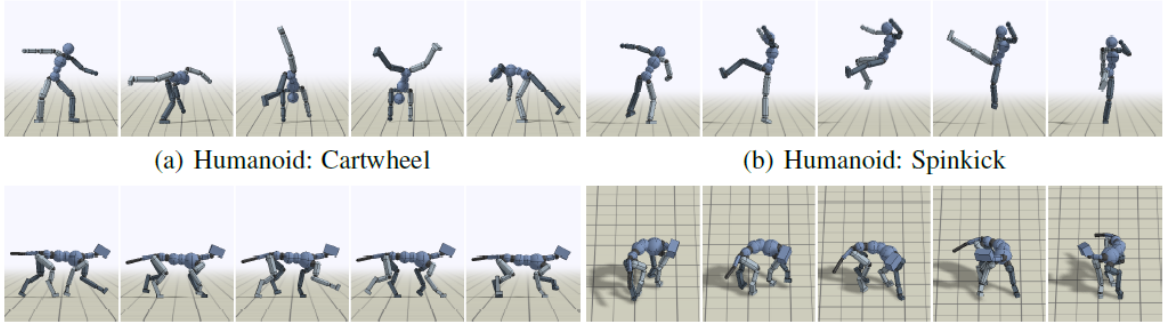
The deadly triad

- RL algorithms combining function approximators, off-policy learning and bootstrap can run instable
- Studied in DQN (does not occur much), but also true for continuous actions
- Three options: no function approximators (tabular RL), being on-policy (TRPO, PPO, ACKTR, A3C), no bootstrap (AWR)
- Being on-policy, TRPO, PPO are not sample efficient
- Without bootstrap, AWR offers an original solution

Richard Sutton. Introduction to reinforcement learning with function approximation. In *Tutorial at NIPS*, 2015

Hado van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep Reinforcement Learning and the Deadly Triad. *arXiv preprint arXiv:1812.02648*, 2018

AWR



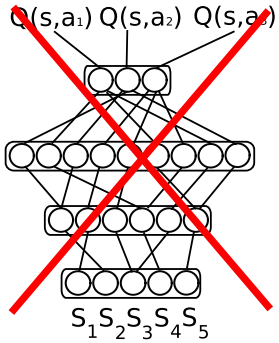
(a) Humanoid: Cartwheel (b) Humanoid: Spinkick

(c) Dog: Trot (d) Dog: Turn

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019

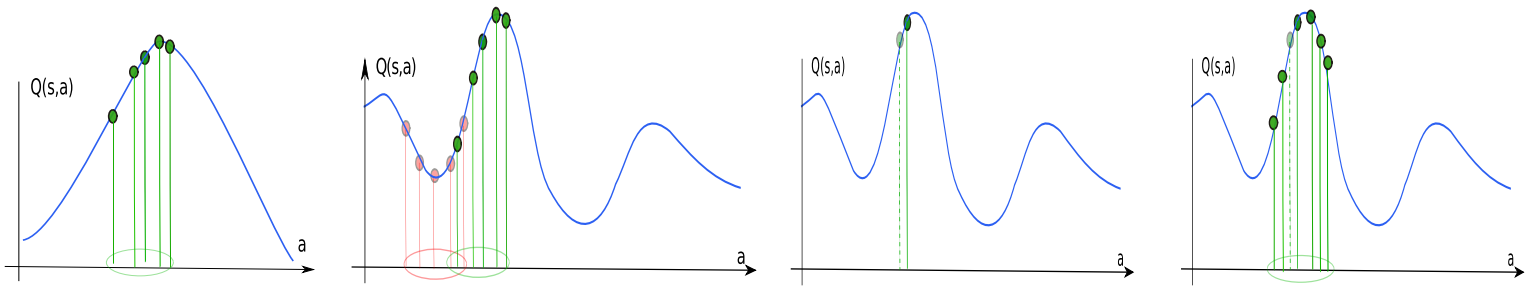
- Approximates an actor and a critic using a standard regression routine
- More precisely, learning the actor uses advantage weighted regression
- Using regression from samples makes it more off-policy
- But still a natural gradient constraint, so not truly off-policy
- Good at learning by imitation (start from expert samples) and even offline learning (just one iteration)
- See my dedicated video: `https://www.youtube.com/watch?v=ivONO2X-MHk`

Finding a max in continuous action space



$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a \in A} Q(s_{t+1}, a) - Q(s_t, a_t)]$$

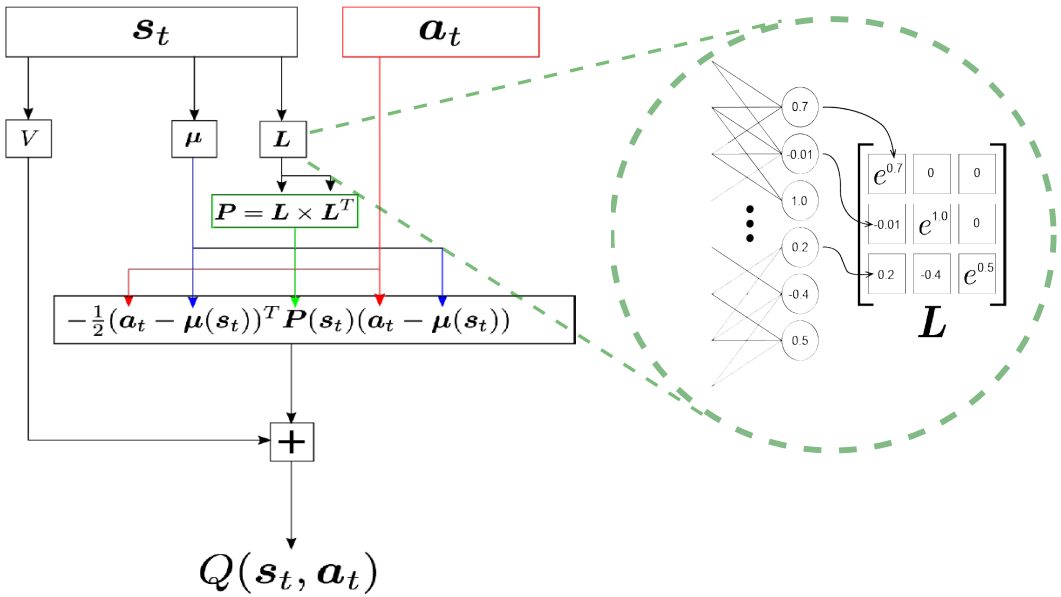
- Two things become too hard:
  - Computing $\max_{a \in A} Q(s_{t+1}, a)$ in the update rule
  - Selecting actions by finding $\max_{a \in A} Q(s, a)$ (as in Q-LEARNING)



- Four 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
  4. SAC: train a stochastic sampler of the best action

NAF

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever and Sergey Levine. (2016) Continuous deep Q-learning with model-based acceleration, *arXiv preprint arXiv:1603.00748*



- The μ network is the actor
- Outperforms DDPG on some benchmarks
- Other tricks in the paper: use iLQG for model-based acceleration
- Used in real robotics settings with some success

Shixiang Gu, Ethan Holly, Timothy Lillicrap and Sergey Levine (2016) Deep reinforcement learning for robotic manipulation. *arXiv preprint arXiv:1610.00633*

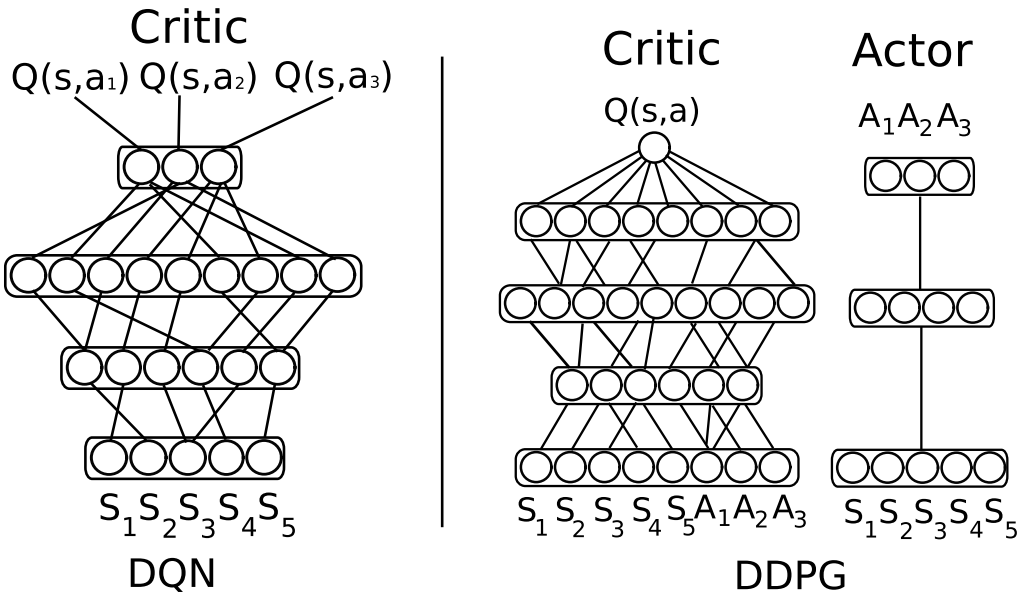
QT-Opt

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

- Used for vision-based grasping, trained on 580K real grasps
- Focus on Section 4.2: how to deal with continuous actions
- To approximate $\operatorname{argmax}_a Q(s, a)$, uses the CEM algorithm
- Two iterations taking the best 6 samples over 64 (so 128 actions)

Training the critic and the actor

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*

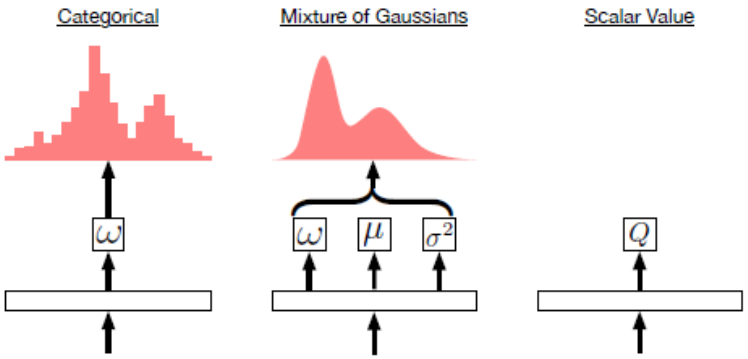


- $\delta_t = r_t + \gamma Q_\theta(s_{t+1}, \pi(s_{t+1})) - Q_\theta(s_t, a_t)$
- If $\pi(s_{t+1})$ close to max, close to Q-learning
- $\nabla_a Q_\theta(s, a)$ is used as error signal to update the actor weights.
- Comes from NFQCA

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

C51

Marc G Bellemare, Will Dabney, and Rémi Munos (2017) A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*



- Distributional policy gradient
- Uses a distribution of rewards and a deterministic policy
- The distribution is richer than just Gaussian
- Interesting theoretical puzzles and concepts (Wasserstein metrics)...
- Biological plausibility of distribution of rewards (population encoding)

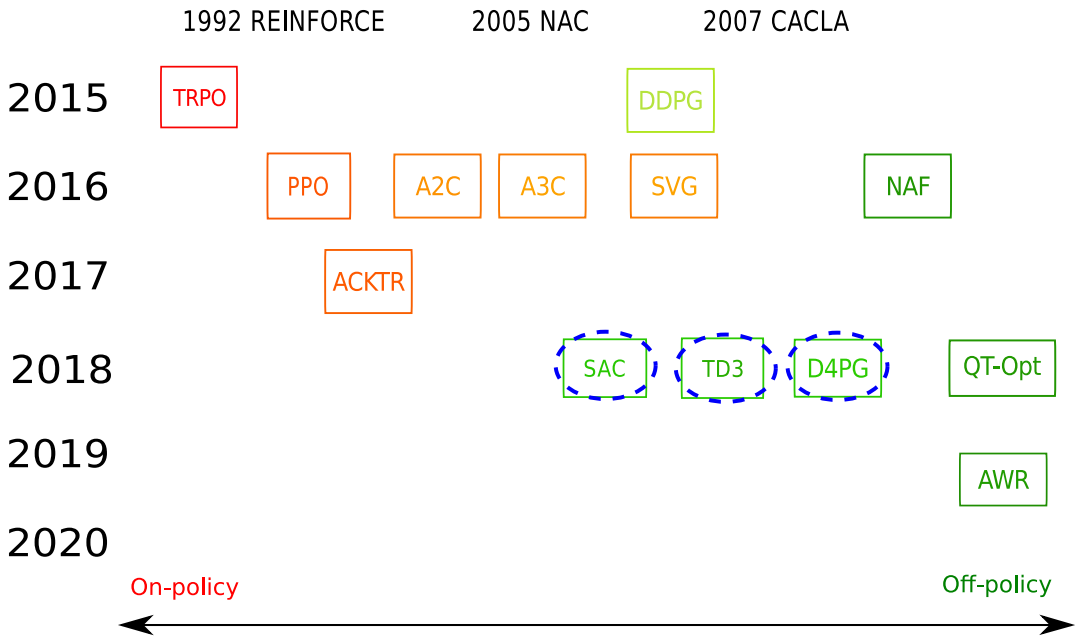
Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pp. 1–5, 2020.

D4PG

Gabriel Barth-maron, Matthew Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy P. Lillicrap (2018) Distributional policy gradient. In *ICLR* (pp. 1–16).

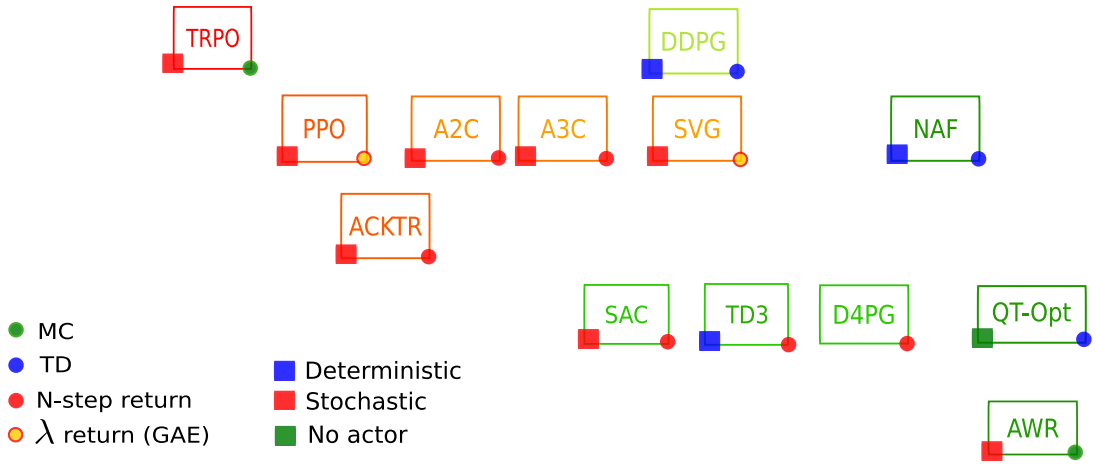
- Combines C51, DDPG, Prioritized Experience Replay and N-step return
- Ouperforms PPO
- The N-step return is shown to be critical

Ranking per year



- It all started in 2015, 2018 was a major year
- Since 2019, the focus is more on other questions: multitask RL, intrinsic motivations, exploration, meta-RL...

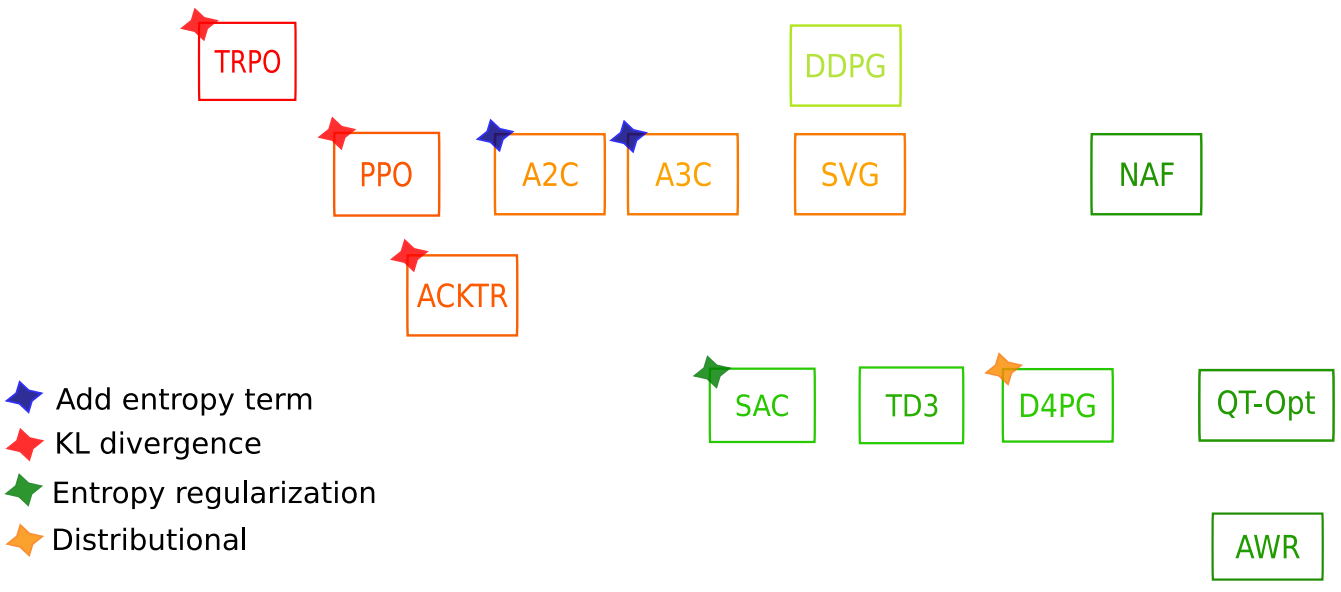
N-step return or not?



- The N-step return is a common ingredient of many SOTA algorithms
- May render an off-policy algorithm a bit more on-policy
- A large conclusive study is missing
- "Counterintuitively we show that theoretically ungrounded, uncorrected n-step returns are uniquely beneficial while other techniques confer limited benefit for sifting through larger memory."

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. (2020) Revisiting fundamentals of experience replay. *arXiv preprint arXiv:2007.06700*

Stochastic policies, KL and Entropy regularization



- Stochastic policies are good for exploration,
- Likelihood of actions changes more smoothly than with deterministic policies
- KL constraint mostly prevents premature convergence to determinism
- So does entropy regularization

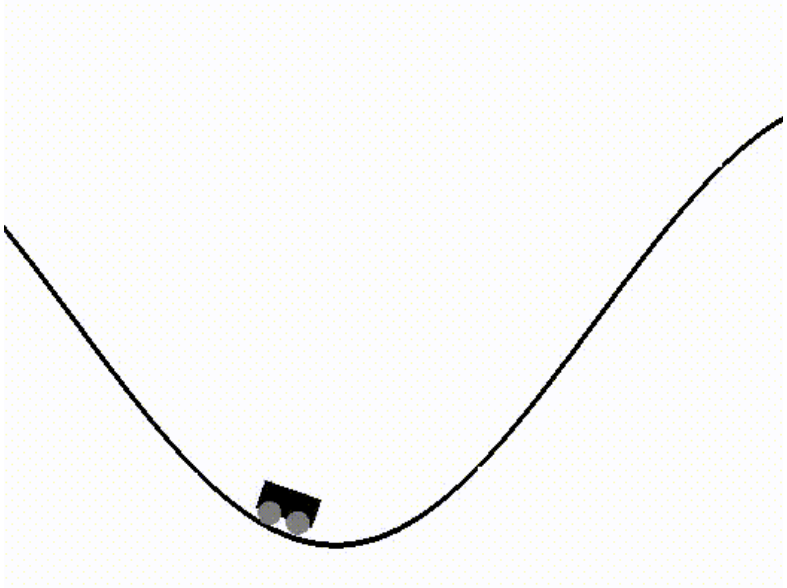
Status

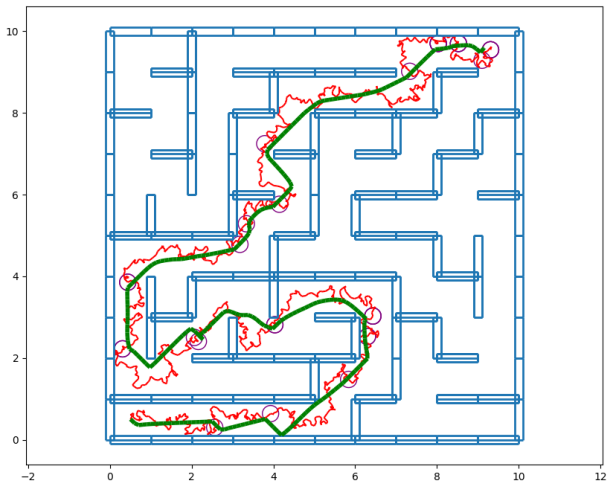
- Large computational ressources are necessary
- Good engineers help a lot
- Grad student descent
- Big actors are ruling the game: Deepmind, OpenAI, Berkeley, Microsoft, FAIR...
- Focus more on performance than on understanding
- Deep RL that matters: instabilities, hard to compare, sensitivity to hyper-parameters
- Empirical comparisons based mostly on openAI, mujoco, Deepmind Control Suite
- The reproducibility crisis, and challenge
- Lack of controlled experiments
- Still fast performance progress, but progress is now more in exploration, multitask learning, curriculum learning, etc.


Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger (2017) Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*


Artemij Amiranashvili, Alexey Dosovitskiy, Vladlen Koltun, and Thomas Brox (2018) TD or not TD: Analyzing the role of temporal differencing in deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Exploration



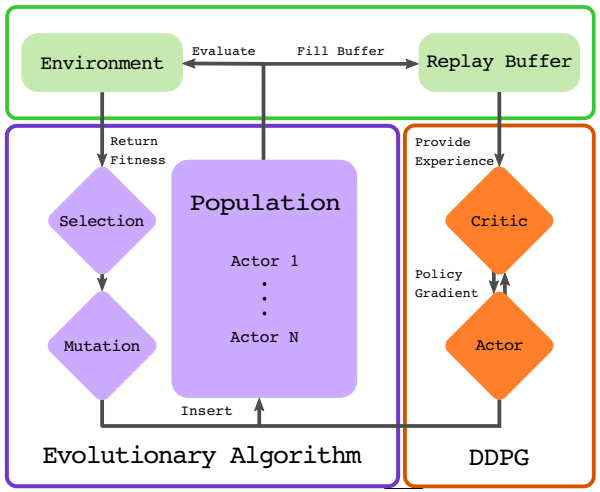


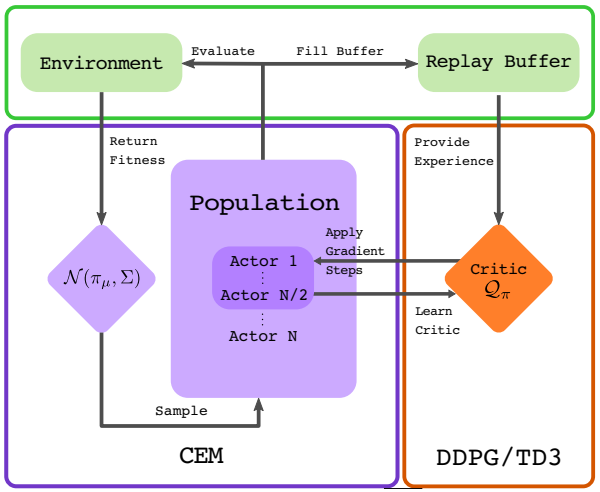
- Exploration is central to RL
- Stochastic policies, KL and Entropy regularization are a way
- But combining deterministic policies with dedicated exploration mechanisms may be preferred
- A class dedicated to exploration should come some day

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer (2018) GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms. *arXiv preprint arXiv:1802.05054*

Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud (2020) PBCS: Efficient exploration and exploitation using a synergy between learning and motion planning. *arXiv preprint arXiv:2004.11667*

Opening: Combining Evo and Deep RL



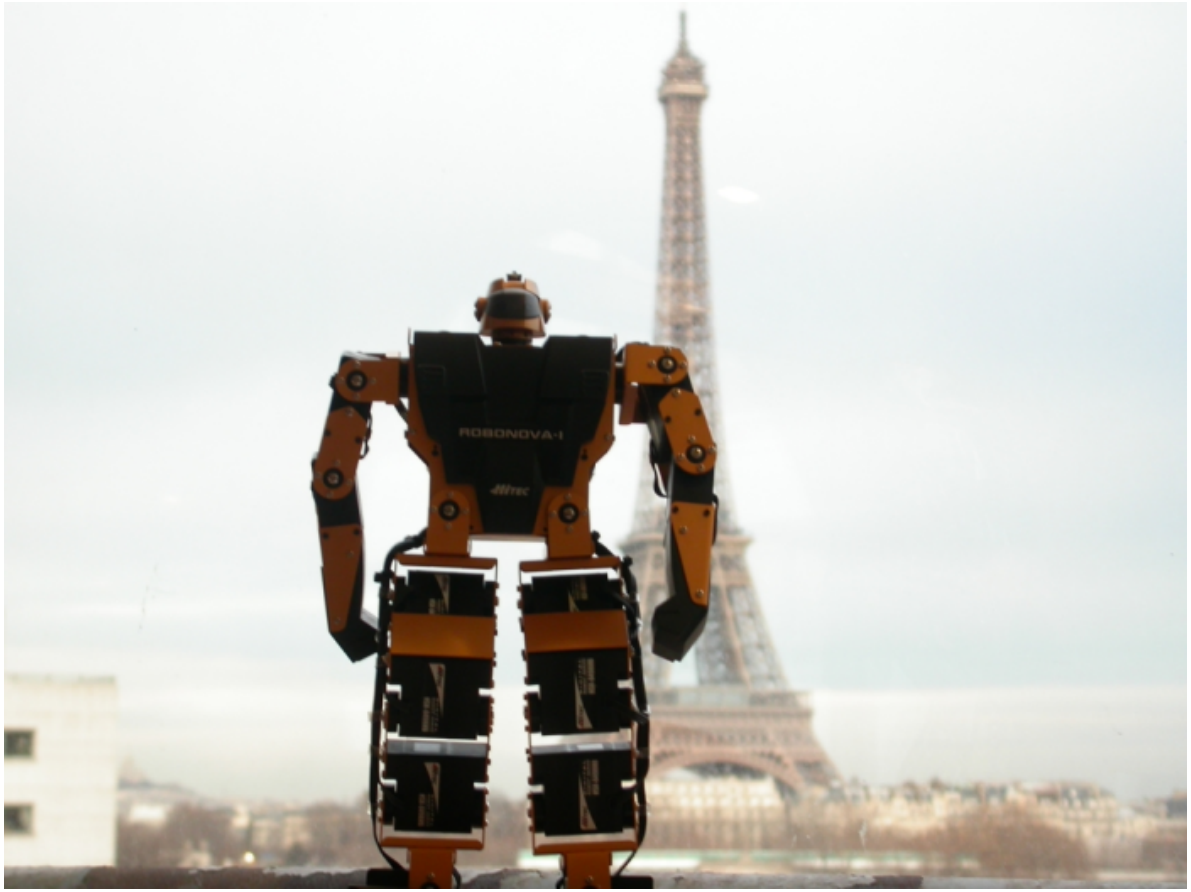


- Combining evolutionary methods and deep RL is an emerging domain
- A lot related to the exploration issue
- In summer 2020, more than 20 new algorithms
- New tendency: look for diversity

Shauharda Khadka and Kagan Tumer (2018a) Evolution-guided policy gradient in reinforcement learning. In *Neural Information Processing Systems*

Aloïs Pourchot and Olivier Sigaud (2018) CEM-RL: Combining evolutionary and gradient-based methods for policy search. *arXiv preprint arXiv:1810.01222* (ICLR 2019)

Any question?



Send mail to: Olivier.Sigaud@upmc.fr


Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans.
Understanding the impact of entropy on policy optimization.
In *International Conference on Machine Learning*, pp. 151–160, 2019.


Artemij Amiranashvili, Alexey Dosovitskiy, Vladlen Koltun, and Thomas Brox.
Td or not td: Analyzing the role of temporal differencing in deep reinforcement learning.
In *International Conference on Learning Representations (ICLR)*, 2018.


Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al.
What matters in on-policy reinforcement learning? a large-scale empirical study.
arXiv preprint arXiv:2006.05990, 2020.


Gabriel Barth-maron, Matthew Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy P. Lillicrap.
Distributional policy gradient.
In *ICLR*, pp. 1–16, 2018.


Marc G Bellemare, Will Dabney, and Rémi Munos.
A distributional perspective on reinforcement learning.
arXiv preprint arXiv:1707.06887, 2017.


Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer.
GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms.
arXiv preprint arXiv:1802.05054, 2018.


Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick.
A distributional code for value in dopamine-based reinforcement learning.
Nature, pp. 1–5, 2020.


William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney.
Revisiting fundamentals of experience replay.
arXiv preprint arXiv:2007.06700, 2020.


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


Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine.
Deep reinforcement learning for robotic manipulation.
arXiv preprint arXiv:1610.00633, 2016a.


Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine.
Q-prop: Sample-efficient policy gradient with an off-policy critic.
arXiv preprint arXiv:1611.02247, 2016b.


Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine.
Continuous deep q-learning with model-based acceleration.
arXiv preprint arXiv:1603.00748, 2016c.


Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, Bernhard Schölkopf, and Sergey Levine.
Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning.
arXiv preprint arXiv:1706.00387, 2017.


Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine.
Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.
arXiv preprint arXiv:1801.01290, 2018a.


Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al.
Soft actor-critic algorithms and applications.
arXiv preprint arXiv:1812.05905, 2018b.


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


Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger.
Deep reinforcement learning that matters.
arXiv preprint arXiv:1709.06560, 2017.


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.


Shauharda Khadka and Kagan Tumer.
Evolution-guided policy gradient in reinforcement learning.
In *Neural Information Processing Systems*, 2018.


Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra.
Continuous control with deep reinforcement learning.
arXiv preprint arXiv:1509.02971, 2015.


Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud.
PBCS: Efficient exploration and exploitation using a synergy between reinforcement learning and motion planning.
arXiv preprint arXiv:2004.11667, 2020.


Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu.
Asynchronous methods for deep reinforcement learning.
arXiv preprint arXiv:1602.01783, 2016.


Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare.
Safe and efficient off-policy reinforcement learning.
In *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2016.


John Paisley, David Blei, and Michael Jordan.
Variational bayesian inference with stochastic search.
arXiv preprint arXiv:1206.6430, 2012.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine.

Advantage-weighted regression: Simple and scalable off-policy reinforcement learning.
arXiv preprint arXiv:1910.00177, 2019.

Aloïs Pourchot and Olivier Sigaud.

CEM-RL: Combining evolutionary and gradient-based methods for policy search.
arXiv preprint arXiv:1810.01222, 2018.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel.

Trust region policy optimization.
CoRR, abs/1502.05477, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov.

Proximal policy optimization algorithms.
arXiv preprint arXiv:1707.06347, 2017.

Richard Sutton.

Introduction to reinforcement learning with function approximation.
In *Tutorial at NIPS*, 2015.

Hado Van Hasselt and Marco A. Wiering.

Reinforcement learning in continuous action spaces.
In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 272–279, 2007.

Hado van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil.

Deep Reinforcement Learning and the Deadly Triad.
arXiv preprint arXiv:1812.02648, 2018.
URL `http://arxiv.org/abs/1812.02648`.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas.

Dueling network architectures for deep reinforcement learning.
arXiv preprint arXiv:1511.06581, 2015.

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.

Yuhuai Wu, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba.

Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation.
arXiv preprint arXiv:1708.05144, 2017.