A Survey of the Actor-Critic Family

Benjamin Möckl
University of Augsburg, Germany

What is Actor-Critic RL?

Actor-critic methods try to learn the **policy function** (the actor) directly, and at the same time learn the **value-function** (the critic) to assess the quality of the policy and improve it. This enables off-policy learning, improves sample-efficiency and reduces variance compared to pure policy-gradient methods.

Actor-Critic is a umbrella term for RL algorithms, which combine learning a policy and learning a value-function or q-function in a meaningful way. Prominent variants are **A2C**, where the value-function is used as a baseline in the policy-gradient loss, and **DPG**, where the policy is trained on maximising a learned Q-function in encountered states. The basic framework can be seen in figure 1.

Algorithms like AlphaGo, where the policy is trained completely separate from the value function, but later combined with a Monte-Carlo-tree-search, might also be considered Actor-Critic, despite their training algorithms are not.

Actor-Critic

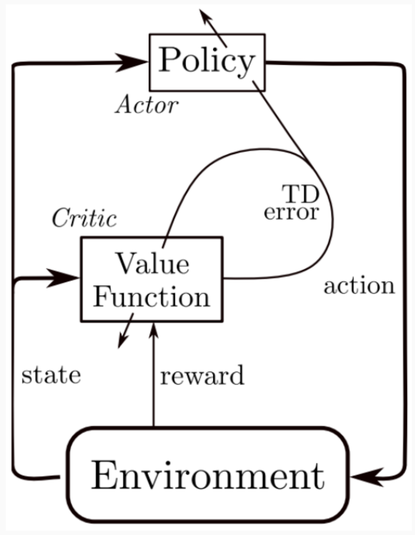

Fig. 1: Overview over the Actor-Critic framework (Sutton&Barto 2018).

History of Actor-Critic

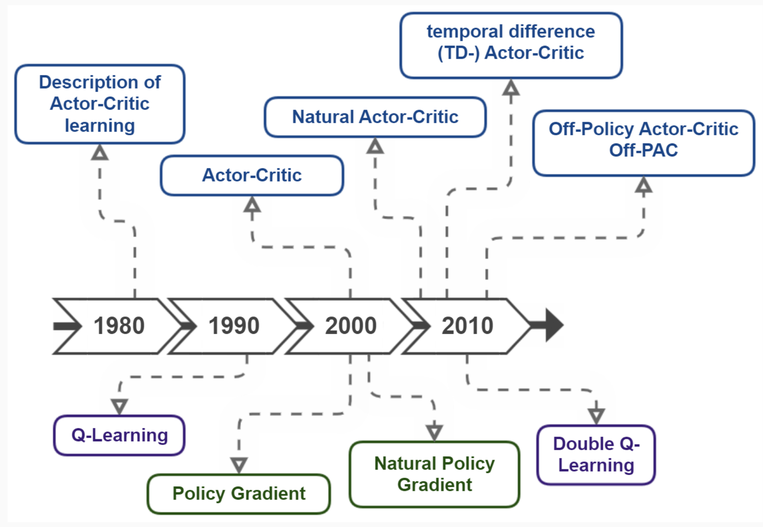

Fig. 2: History of Actor-Critic methods predating deep-learning. Colour-coding see caption of figure 3.

Milestones of RL

- **2013: DQN** Use of deep neural networks as function approximators in Q-Learning.
- **2014: TRPO** Robust learning of 2D bipedal locomotion.
- **2015: DQN** Human-level control on ATARI games.
- **2016: TRPO+GAE** Robust learning of 3D bipedal locomotion.
- **2016: AlphaGo** Beating the human European Go champion by 5 games to 0.
- **2017: AlphaZero** Surpassing AlphaGo without the use of human knowledge.
- **2017: Rainbow-DQN** Combining different improvements of DQN successfully.
- **2019: OpenAI Five** Winning against the world-champion team in DotA2.
- **2019: AlphaStar** Reaching Grandmaster level in the full game of StarCraft II.
- **2021: DreamerV2** Human-level performance on ATARI games by learning behaviors only inside a separately trained world model.

History of Actor-Critic 2013 - 2022

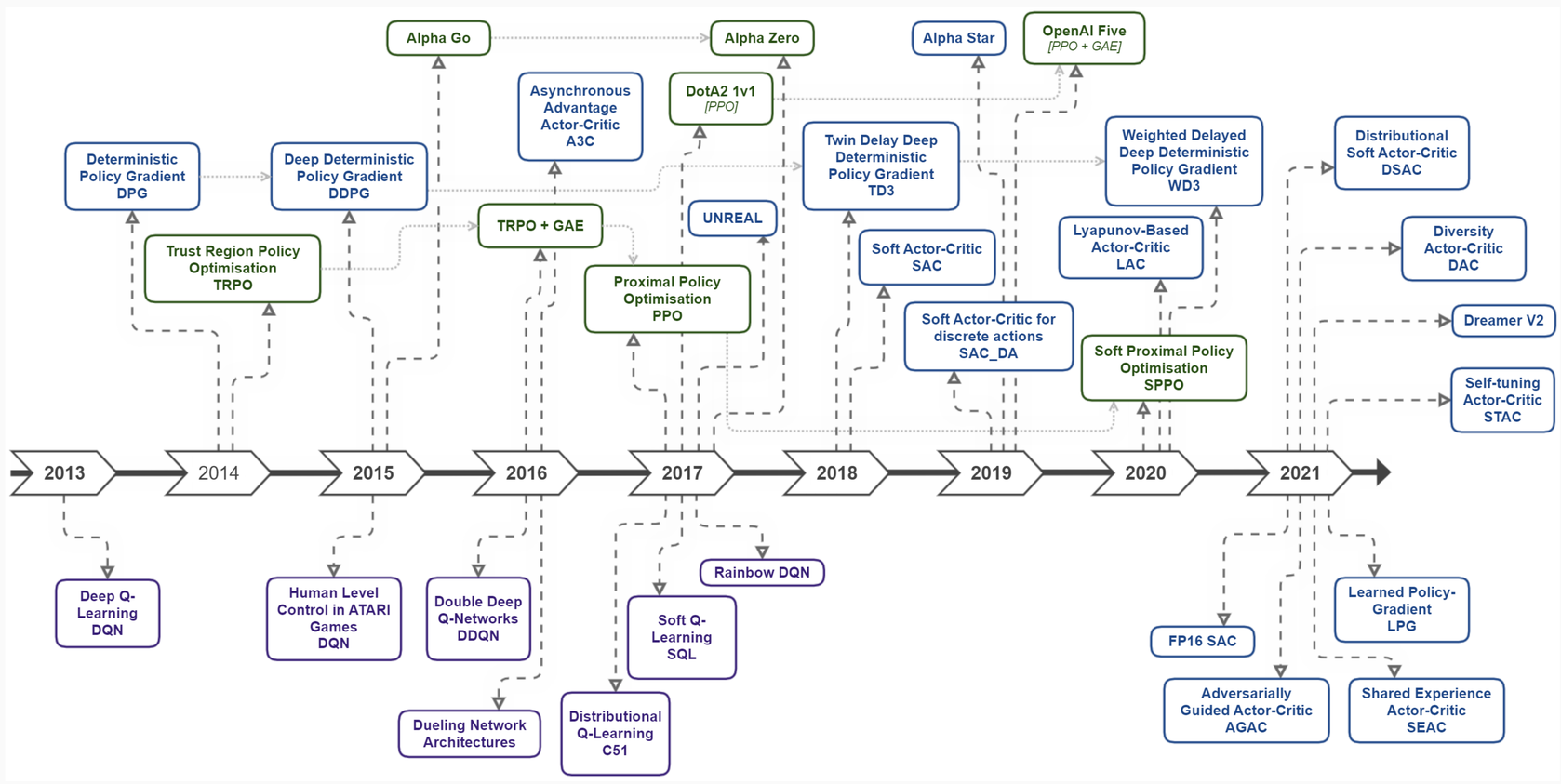

Fig. 3: purple: Q-Learning based methods; green: Policy-Learning based methods; blue: Actor-Critic based methods. Every bubble is related to a research paper. The list is focused on biggest breakthroughs throughout time related to current day actor-critic methods, non-exhaustive.

State-of-the-Art 2021

DSAC: Distributional Soft Actor-Critic integrates learning the **distribution of rewards** instead of the expected sum of future rewards in its critic and adds a **max-entropy term** to its objective. The goal of this was to reduce overestimation bias of learned Q-values. This resulted in more stable learning and a new state-of-the-art on the *MuJoCo* benchmark.

WD3: Weighted Delayed Deep Deterministic Policy Gradient uses **two critics** and a new parameter β, that **weights their influence**, regulating between overestimation and underestimation, reducing estimation error and therefore increasing performance. *WD3* reaches a new state-of-the-art performance on the tested *MuJoCo* benchmarks.

DAC: Diversity Actor-Critic builds on top of *SAC* by **extending the entropy maximisation** term to a weighted sum of the current action distribution and the action distribution from the replay-buffer.
The algorithm consistently and significantly outperforms *SAC* on sparse and delayed *MuJoCo* benchmarks.

DreamerV2 is a **model-based** Actor-Critic algorithm, that learns behaviours **purely from predictions** in the compact latent space of its powerful world model. It is the first agent that achieves human-level performance on *ATARI* games while training its policy only inside its world model. It significantly outperforms the top single-GPU agents Rainbow-DQN and IQN.

LAC: Lyapunov Actor-Critic uses a *Lyapunov* critic function to **guarantee closed-loop stability**. Stability is a very important property of any control system because of its close relation to safety, robustness and reliability in robotic systems. The algorithm was evaluated on *MuJoCo*, and is significantly more robust while performing on par with *SAC* and *SPPO* methods.

STAC: Self-Tuning Actor-Critic uses metagradients to **self-tune all differentiable hyperparameters** in an actor-critic loss function, online and within a single lifetime. Self-tuning increased the performance of the original algorithm, while being computationally efficient and robust to its hyperparameters.

FP16 SAC successfully adapted **low precision training** (FP16) to *SAC* with a set of easy-to-implement modifications. Their evaluation shows the low precision *SAC* variant *FP16 SAC* to **match the performance** of its full precision counterpart, with lower memory and compute costs.

LPG: Learned Policy-Gradient is the result of an approach at meta-learning for discovering new reinforcement learning algorithms. The algorithm was shown to **discover concepts** like value functions and bootstrapping. It generalizes effectively to complex *ATARI* games and matches the performance of *A2C* in some of them.

AGAC: Adversarially Guided Actor-Critic introduces a **third component** to actor and critic: the **adversary**. It tries to mimic the actors action distribution, while the actor tries to differentiate itself from it. This induces search for novel strategies, similar to *DAC*. AGAC outperforms the previous state-of-the-art algorithms on various hard-exploration benchmarks.

SEAC: Shared Experience Actor-Critic applies **experience sharing** between multiple agents to the actor-critic framework. It updates the actor and critic parameters of an agent by combining gradients computed on the agents experience with weighted ones computed on other agents experiences. It achieves state-of-the-art performance in sparse multi-agent environments.