Policy Gradient

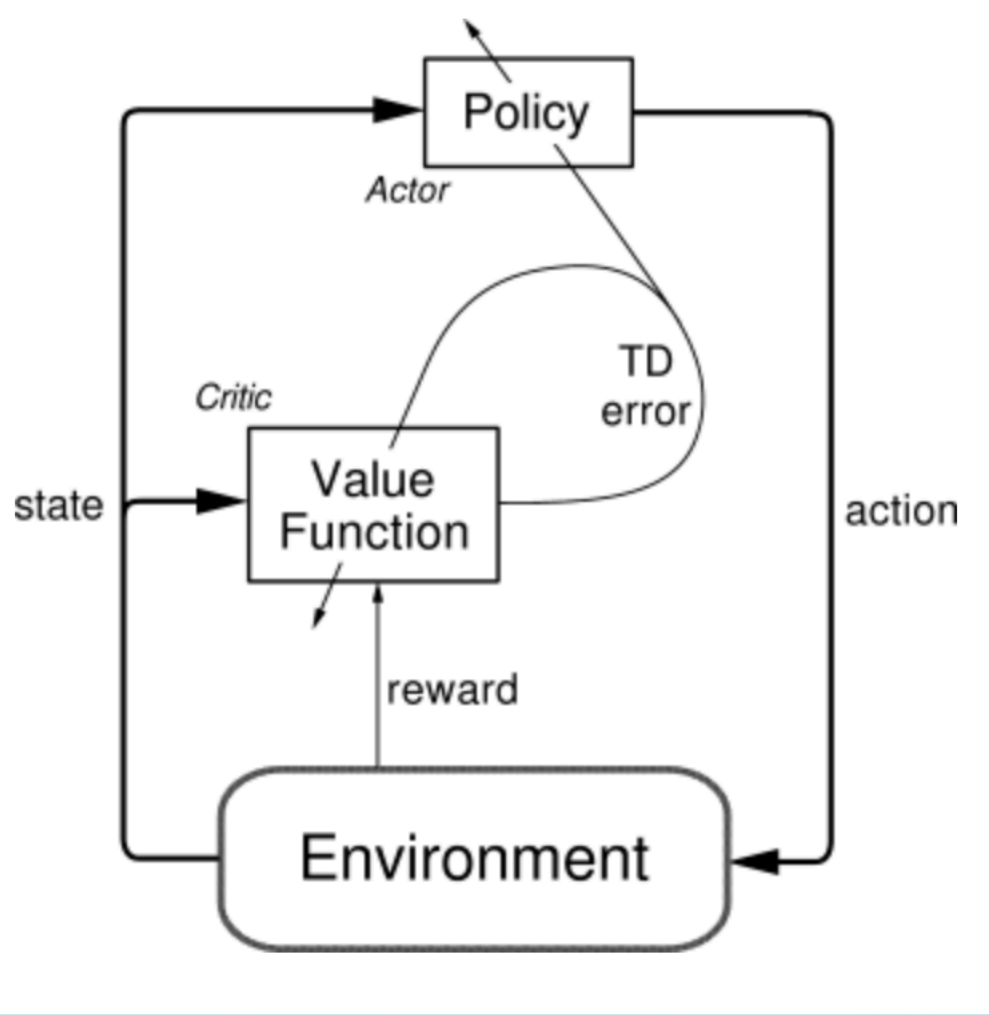
- David Silver's Slides presented by Jonathan Lorraine
- Images in this powerpoint provided by:

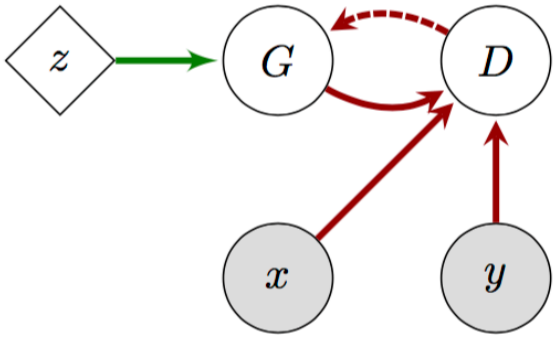
David Pfau, Oriol Vinyals. *Generative Adversarial Networks and Actor-Critic Methods.*
arXiv:1610.01945

Videos

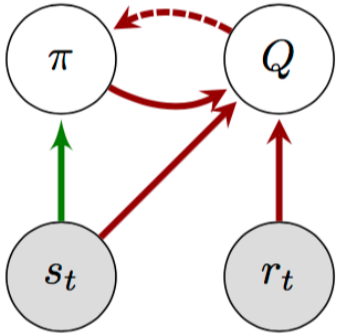
Degris – Pendulum with pdf

Local Mininimum Convergence


Policy
Actor
Critic
TD
error
Value
Function
state
action
reward
Environment



(a) Generative Adversarial Networks [4]



(b) Deterministic Policy Gradient [7] / SVG(0) [8] / Neurally-Fitted Q-learning with Continuous Actions [9]

Figure 1: Information structure of GANs and AC methods. Empty circles represent models with a distinct loss function. Filled circles represent information from the environment. Diamonds represent fixed functions, both deterministic and stochastic. Solid lines represent the flow of information while dotted lines represent the flow of gradients used by another model. Paths which are analogous between the two models are highlighted in red. The dependence of Q on future states and the dependence of future states on π are omitted for clarity.

Both GANs and AC can be seen as bilevel or two-time-scale optimization problems, where one model is optimized with respect to the optimum of another model:

$$x^* = \arg\min_{x \in \mathcal{X}} F(x, y^*(x)) \tag{1}$$

$$y^*(x) = \arg\min_{y \in \mathcal{Y}} f(x, y) \tag{2}$$

$$Q^{\pi}(s,a) = \mathbb{E}_{s_{t+k}\sim\mathcal{P}, r_{t+k}\sim\mathcal{R}, a_{t+k}\sim\pi}\left[\sum_{k=1}^{\infty}\gamma^{k} r_{t+k}\middle| s_t = s, a_t = a\right] \tag{6}$$

and learn a policy that is optimal for that value function:

$$\pi^{*} = \arg\max_{\pi} \mathbb{E}_{s_0\sim p_0, a_0\sim\pi}[Q^{\pi}(s_0,a_0)] \tag{7}$$

We can express Q^{π} as the solution to a minimization problem:

$$Q^{\pi} = \arg\min_{Q} \mathbb{E}_{s_t,a_t\sim\pi}[\mathcal{D}(\mathbb{E}_{s_{t+1},r_t,a_{t+1}}[r_t + \gamma Q(s_{t+1},a_{t+1})]||Q(s_t,a_t))] \tag{8}$$

Where $\mathcal{D}(\cdot||\cdot)$ is any divergence that is positive except when the two are equal. Now the actor-critic problem can be expressed as a bilevel optimization problem as well:

$$F(Q,\pi) = \mathbb{E}_{s_t,a_t\sim\pi}[\mathcal{D}(\mathbb{E}_{s_{t+1},r_t,a_{t+1}}[r_t + \gamma Q(s_{t+1},a_{t+1})]||Q(s_t,a_t))] \tag{9}$$

$$f(Q,\pi) = -\mathbb{E}_{s_0\sim p_0, a_0\sim\pi}[Q^{\pi}(s_0,a_0)] \tag{10}$$