CHAPTER 1

ADVANCED ALGORITHMS

Temporal-difference (TD) learning control methods can be grouped into on-policy and off-policy methods. (Control methods also solve for prediction problems) A third class of these TD control methods are actor-critic.

1.1 Overview

The problem with the value-function approach to finding policies are those policies are deterministic, whereas optimal policies are often stochastic. The second drawback is a small change in the value of an action can change which action is selected for the given state. [3]

Actor-critic refers to a family of algorithms. First introduced by Barto et al. [1]

Example of multiple continuous action: [2]

1.2 Policy Gradient Methods

A policy is the probability that action a is taken at time t when the agent is in state s at time t with parameter θ. In other words,

$$\pi(a|s,\boldsymbol{\theta}) = \Pr\{A_t = a|S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}$$

If the action space is discrete and not too large, an exponential soft-max distribution can be used in action selection,

$$\pi(a|s,\boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}},$$

where

$$h(s,a,\boldsymbol{\theta}) = \boldsymbol{\theta}^{\top}\mathrm{x}(s,a),$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and $\mathrm{x}(s,a) \in \mathbb{R}^{d'}$ is the feature vector.

Algorithm 1: Actor-critic for episodic problems

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathrm{w})$
Parameter(s): step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathrm{w}} > 0$

Initialise policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathrm{w} \in \mathbb{R}^{d}$ (e.g, to 0) ;
for *each episode* **do**
  Initialise S (first state of episode);
  $I \leftarrow 1$;
  for *each step of episode or until S is terminal* **do**
    $A \sim \pi(.|S,\boldsymbol{\theta})$;
    Take action A, observe S', R;
    $\delta \leftarrow R + \gamma\hat{v}(S',\mathrm{w}) - \hat{v}(S,\mathrm{w})$ (if S' is terminal, $\hat{v}(S',\mathrm{w}) \doteq 0$) ;
    $\mathrm{w} \leftarrow \mathrm{w} + \alpha^{\mathrm{w}}\delta\nabla\hat{v}(S,\mathrm{w})$;
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I\delta\nabla\pi(A|S,\boldsymbol{\theta})$;
    $I \leftarrow \gamma I$;
    $S \leftarrow S'$;
  end
end

Note that in actor-critic with traces with linear function approximation $\nabla\hat{v}(S,\mathrm{w})$ reduces to the feature vector $\mathrm{x}(S,a)$

Algorithm 2: Actor-critic with eligibility traces for episodic problems

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathrm{w})$
Parameter(s): step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathrm{w}} > 0$
Parameter(s): trace decay rates $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\lambda^{\mathrm{w}} \in [0,1]$
Initialise policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathrm{w} \in \mathbb{R}^{d}$ (e.g, to 0) ;
for *each episode* **do**
  Initialise S (first state of episode);
  $\mathrm{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ (d'-component eligibility trace vector) ;
  $\mathrm{z}^{\mathrm{w}} \leftarrow \mathbf{0}$ (d-component eligibility trace vector) ;
  $I \leftarrow 1$;
  for *each step of episode or until S is terminal* **do**
    $A \sim \pi(.|S,\boldsymbol{\theta})$;
    Take action A, observe S', R;
    $\delta \leftarrow R + \gamma\hat{v}(S',\mathrm{w}) - \hat{v}(S,\mathrm{w})$ (if S' is terminal, $\hat{v}(S',\mathrm{w}) \doteq 0$) ;
    $\mathrm{z}^{\mathrm{w}} \leftarrow \gamma\lambda^{\mathrm{w}}\mathrm{z}^{\mathrm{w}} + \nabla\hat{v}(S,\mathrm{w})$;
    $\mathrm{z}^{\boldsymbol{\theta}} \leftarrow \gamma\lambda^{\boldsymbol{\theta}}\mathrm{z}^{\boldsymbol{\theta}} + I\nabla \ln \pi(A|S,\boldsymbol{\theta})$;
    $\mathrm{w} \leftarrow \mathrm{w} + \alpha^{\mathrm{w}}\delta\mathrm{z}^{\mathrm{w}}$;
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\delta\mathrm{z}^{\boldsymbol{\theta}}$;
    $I \leftarrow \gamma I$;
    $S \leftarrow S'$;
  end
end