Algorithms in Reinforcement Learning

Guannan Hu

March 14, 2018

Algorithm 1 Actor-Critic with Eligibility Traces (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\hat{v}(s, \mathbf{w})$
Input: a differentiable state-value function parameterization $\hat{v}(s, \mathbf{w})$
Algorithm parameters: trace-decay rate $\lambda^{\boldsymbol{\theta}} \in [0, 1], \lambda^{\mathbf{w}} \in [0, 1]$; step sizes $\alpha^{\boldsymbol{\theta}} > 0, \alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d^{'}}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

loop forever (for each episode):
  initialize S (first state of episode)
  $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ ($d^{'}$-component eligibility trace vector)
  $\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ ($d^{'}$-component eligibility trace vector)
  $\mathbf{I} \leftarrow \mathbf{1}$
  loop while S is not terminal (for each time step):
    $A \in \pi(\cdot|S, \boldsymbol{\theta})$
    Take action A, observe $S^{'}, R$
    $\delta \leftarrow \gamma\hat{v}(S^{'}, \mathbf{w}) - \hat{v}(S, \mathbf{w})$ (if $S^{'}$ is terminal, then $\hat{v}(S^{'}, \mathbf{w})$)
    $\mathbf{z}^{\mathbf{w}} \leftarrow \gamma\lambda^{\mathbf{w}}\mathbf{z}^{\mathbf{w}} + I\nabla\hat{v}(s, \mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \gamma\lambda^{\boldsymbol{\theta}}\mathbf{z}^{\boldsymbol{\theta}} + I\nabla\hat{v}(s, \boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}\delta\mathbf{z}^{\mathbf{w}}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\theta}\delta\mathbf{z}^{\boldsymbol{\theta}}$
    $\mathbf{I} \leftarrow \gamma\mathbf{I}$
    $S \leftarrow S^{'}$
  end loop
end loop

Algorithm 2 Actor-Critic with Eligibility Traces (continuing), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s, \mathbf{w})$
Algorithm parameters: trace-decay rates $\lambda^{\boldsymbol{\theta}} \in [0,1], \lambda^{\mathbf{w}} \in [0,1]$; step sizes $\alpha^{\boldsymbol{\theta}} > 0, \alpha^{\mathbf{w}} > 0, \eta > 0$
Initialize $\bar{R} \in \mathbb{R}$ (e.g., to 0)
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$ (e.g., to $\mathbf{0}$) $\doteq 0$
loop forever (for each time step):
    $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
    Take Action A, observe S', R
    $\delta \leftarrow R - \bar{R} + \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$
    $\bar{R} \leftarrow \bar{R} + \eta\delta$
    $\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}}\mathbf{z}^{\mathbf{w}} + \nabla\hat{v}(S, \mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}}\mathbf{z}^{\boldsymbol{\theta}} + \nabla \ln \pi(A|S, \boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\delta\mathbf{z}^{\mathbf{w}}$
    $\mathbf{w} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\delta\mathbf{z}^{\boldsymbol{\theta}}$
    $S \leftarrow S'$
end loop