5.4 Optimistic Policy Iteration

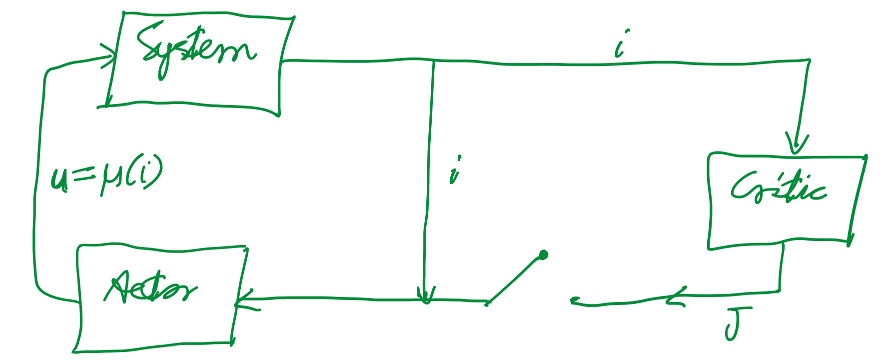
Friday, February 12, 2021 10:53 AM

The algorithm fixes a policy μ, evaluates J^μ, and then perform a policy update.

Updates are carried out using Actor-Critic system wherein actor uses policy μ to control the system and critic computes J^μ.

Actions are chosen to minimize,

$$\sum_j p_{ij}(u)\left(g(i,u,j) + J(j)\right).$$



We can perform a policy update corresponding to every update of policy evaluation algorithm (critic). Thus, we perform updates at each simulation step. These type of methods are called optimistic policy iteration.

<u>Visualizing Policy Iteration</u> –

We say that μ is greedy w.r.to J if μ attains the minimum,

$$\sum_j p_{ij}(\mu(i))\left(g(i,\mu(i),j) + \alpha J(j)\right) = \min_{u \in U(i)} \sum_j p_{ij}(u)\left(g(i,u,j) + \alpha J(j)\right).$$

$$\Rightarrow T_\mu J = TJ.$$

Let R_μ be the set of vectors J that lead to greedy policy,

$$R_\mu = \{ J \mid \mu \text{ is greedy w.r.to } J \}.$$

$$\Rightarrow J \in R_\mu \text{ iff. } T_\mu J = TJ.$$

$$\Rightarrow \sum_j p_{ij}(\mu(i))\left(g(i,\mu(i),j) + \alpha J(j)\right) \le \sum_j p_{ij}(u)\left(g(i,u,j) + \alpha J(j)\right).$$

Since this is a set of linear inequalities, R_μ is a polyhedron.

<u>Optimistic TD(0)</u>.

Consider the online TD(0) algorithm,

$$J(i) = J(i) + \gamma\left[g(i,\mu(i),j) + \alpha J(j) - J(i)\right]$$

$$= (1-\gamma)J(i) + \gamma\left[g(i,\mu(i),j) + \alpha J(j)\right].$$

This is of the form,

$$J(i) = (1-\gamma)J(i) + \gamma(T_\mu J)(i) + \gamma w.$$

zero mean noise term

With a diminishing stepsize,

$$J(i) = (1-\gamma)J(i) + \gamma(T_\mu J)(i).$$

In the case of optimistic TD(0), policy μ updated at each update is a greedy policy w.r.t. current vector J and satisfies $T_\mu J = TJ$.

$$J(i) = (1-\gamma)J(i) + \gamma(TJ)(i) + \gamma w.$$

<u>Optimistic TD(1)</u>.

Consider the case when TD(1) is used for policy evaluation and $\alpha < 1$ (discounted problem).

The cumulative cost for every state i_k,

$$g(i_k, i_{k+1}) + \alpha g(i_{k+1}, i_{k+2}) + \cdots$$

This yields an unbiased estimate of $J^\mu(i_k)$.

In theory, the synchronous version of TD(1) is guaranteed to converge to J^* with probability 1.

x ———— x ———— x