On-policy Stability of TD(0)

By Mohammad Pezeshki

Value function $V(s)$ approximation: $V : S \to R$

$\in \mathbb{R}^{|S|} \rightarrow$ Huge

Function approximation with $n \ll |S|$ parameters

$$V(s) = \theta^T \phi(s)$$

Feature vector

Parameter vector

$\in \mathbb{R}^n$

TD(0) update rule at time-step $t+1$:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \Big(\underbrace{\underbrace{R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(S_{t+1})}_{\text{TD Target}} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(S_t)}_{\text{TD Error}} \Big) \boldsymbol{\phi}(S_t)$$

What we want to show? **TD(0) with above update rule is convergent.**

On-policy Stability of TD(0)

TD(0) update rule at time-step $t+1$:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \Big(R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(S_{t+1}) - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(S_t) \Big) \boldsymbol{\phi}(S_t)$$

Re-write:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha \Big(\underbrace{R_{t+1}\boldsymbol{\phi}(S_t)}_{\mathbf{b}_t \in \mathbb{R}^n} - \underbrace{\boldsymbol{\phi}(S_t)\left(\boldsymbol{\phi}(S_t) - \gamma \boldsymbol{\phi}(S_{t+1})\right)^\top}_{\mathbf{A}_t \in \mathbb{R}^{n \times n}} \boldsymbol{\theta}_t \Big) \\
&= \boldsymbol{\theta}_t + \alpha(\mathbf{b}_t - \mathbf{A}_t \boldsymbol{\theta}_t) \\
&= (\mathbf{I} - \alpha \mathbf{A}_t)\boldsymbol{\theta}_t + \alpha \mathbf{b}_t .
\end{aligned}$$

$\mathbf{A}_t$ is multiplied in itself in each iteration.

$\mathbf{A}_t < 0 \longrightarrow (\mathbf{I} - \alpha \mathbf{A}_t) > 1 \longrightarrow$ Diverge

$\mathbf{A}_t > 0 \longrightarrow (\mathbf{I} - \alpha \mathbf{A}_t) < 1 \longrightarrow$ Converge

In general, the updates converge whenever $\mathbf{A}_t$ is positive definite.

But $\mathbf{A}_t$ is a random variable $\longrightarrow$ Using its expectation $\lim_{t \to \infty} \mathbb{E}[\mathbf{A}_t]$

On-policy Stability of TD(0)

Some definitions:

The probability of visiting each state: the steady-state distribution:

$$\mathbf{d}_\pi \dashrightarrow [\mathbf{d}_\pi]_s \doteq d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\}$$

The transition probability matrix (from state i to j):

$$[\mathbf{P}_\pi]_{ij} \doteq \sum_a \pi(a|i)p(j|i,a)$$

The special property of $\mathbf{d}_\pi$ is that:

$$\mathbf{P}_\pi^\top \mathbf{d}_\pi = \mathbf{d}_\pi$$

Now, we rewrite the TD(0) update equation in a deterministic way:

$$\mathbf{A} \doteq \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] \quad \mathbf{b} \doteq \lim_{t\to\infty} \mathbb{E}[\mathbf{b}_t]$$

$$\bar{\boldsymbol{\theta}}_{t+1} \doteq \bar{\boldsymbol{\theta}}_t + \alpha(\mathbf{b} - \mathbf{A}\bar{\boldsymbol{\theta}}_t)$$

Means stability

is convergent to a unique fixed point independent of the initial $\bar{\boldsymbol{\theta}}_0$.

On-policy Stability of TD(0)

$$\bar{\boldsymbol{\theta}}_{t+1} \doteq \bar{\boldsymbol{\theta}}_t + \alpha(\mathbf{b} - \mathbf{A}\bar{\boldsymbol{\theta}}_t)$$

is convergent to a unique $\underbrace{\text{fixed point}}_{\bar{\boldsymbol{\theta}} \,=\, \mathbf{A}^{-1}\mathbf{b}}$ independent of the initial $\bar{\boldsymbol{\theta}}_0$.

if and only if

$\mathbf{A}$ has a full set of eigenvalues all of whose real parts are positive.

We prove stability by showing that $\mathbf{A}$ is positive definite

$$\forall \mathbf{y} \quad \mathbf{y}^\top \mathbf{A}\mathbf{y} > 0$$

$$\begin{aligned}\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_\pi\left[\boldsymbol{\phi}(S_t)\left(\boldsymbol{\phi}(S_t) - \gamma\boldsymbol{\phi}(S_{t+1})\right)^\top\right] \\ = \sum_s d_\pi(s)\,\boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\boldsymbol{\phi}(s')\right)^\top \\ = \boldsymbol{\Phi}^\top \mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)\boldsymbol{\Phi}\end{aligned}$$

On-policy Stability of TD(0)

$$\mathbf{A} = \lim_{t\to\infty} \mathbb{E}[\mathbf{A}_t] = \lim_{t\to\infty} \mathbb{E}_\pi\left[\boldsymbol{\phi}(S_t)\left(\boldsymbol{\phi}(S_t) - \gamma\boldsymbol{\phi}(S_{t+1})\right)^\top\right]$$

$$= \sum_s d_\pi(s)\,\boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\boldsymbol{\phi}(s')\right)^\top$$

$$= \boldsymbol{\Phi}^\top \underbrace{\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)}_{\text{Key matrix}}\boldsymbol{\Phi}$$

$|S| \times |S|$

$\mathbf{d}_\pi$ on diagonal

$|S| \times n$

Theorem: $\mathbf{A}$ is positive definite if key matrix is positive definite.

All of its columns sum to a nonnegative number.

On-policy Stability of TD(0)

$\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)$

All of its columns sum to a nonnegative number.

Using the following two theorems

1. Any matrix $\mathbf{M}$ is positive definite if and only if the symmetric matrix $\mathbf{S} = \mathbf{M} + \mathbf{M}^\top$ is positive definite.
2. Any symmetric real matrix S is positive definite if all of its diagonal entries are positive and greater than the sum of the corresponding off-diagonals.

For $\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)$

The diagonals are positive and the off-diagonals are negative.

To show:

Each row sum plus the corresponding column sum is positive.

On-policy Stability of TD(0)

For $\mathbf{D}_\pi(\mathbf{I} - \gamma \mathbf{P}_\pi)$

The diagonals are positive and the off-diagonals are negative.

To show:

Each row sum plus the corresponding column sum is positive.

is positive because $\mathbf{P}_\pi$ is a stochastic matrix and $\gamma < 1$.

Column sum of $\mathbf{M}$: $\mathbf{1}^\top \mathbf{M}$

is non-negative because ...

$$
\begin{aligned}
\mathbf{1}^\top \mathbf{D}_\pi(\mathbf{I} - \gamma \mathbf{P}_\pi) &= \mathbf{d}_\pi^\top (\mathbf{I} - \gamma \mathbf{P}_\pi) \\
&= \mathbf{d}_\pi^\top - \gamma \mathbf{d}_\pi^\top \mathbf{P}_\pi) \\
&= \mathbf{d}_\pi^\top - \gamma \mathbf{d}_\pi^\top \\
&= (1 - \gamma)\mathbf{d}_\pi > 0
\end{aligned}
$$

So $\mathbf{A}$ is positive definite!

On-policy Stability of TD(0)

In summary:

$\mathbf{D}_\pi(\mathbf{I} - \gamma \mathbf{P}_\pi)$: Each row sum plus the corresponding column sum is positive.

$\mathbf{D}_\pi(\mathbf{I} - \gamma \mathbf{P}_\pi)$: Key matrix is positive definite.

$\mathbf{A} = \mathbf{\Phi}^\top \mathbf{D}_\pi(\mathbf{I} - \gamma \mathbf{P}_\pi)\mathbf{\Phi}$: is positive definite.

$\mathbf{A}$: has a full set of eigenvalues all of whose real parts are positive.

$\bar{\boldsymbol{\theta}}_{t+1} \doteq \bar{\boldsymbol{\theta}}_t + \alpha(\mathbf{b} - \mathbf{A}\bar{\boldsymbol{\theta}}_t)$: is convergent to a unique fixed point.