Successor Representation

Hajar Zaid

November 2025

Let x_t be a vector of probabilities representing a current state and let P be the transition matrix where P_{ij} is the probability of moving from state i to j. Given a markov chain

$$x_{t+1} = x_t P$$

where x_t is probabilities of a current state t and x_{t+1} is the vector of probabilities at state $t+1$.

For the transition matrix P and the images of x_t under the first n powers of P, $\{x_t, x_tP, x_tP^2, \dots, x_tP^n\}$, the following properties are true.

1. The L^1 norm of each row equals one, $\|P_i\|_1 = 1$.
2. P has eigenvalues $|\lambda| \leq 1$
3. The power iteration[1] is guaranteed to converge to a fixed distribution if there exists at least one eigenvalue such that $|\lambda| = 1$ and all other eigenvalues satisfy $|\lambda_k| < 1$.
4. If $\lambda_k = e^{i(\theta + \frac{2\pi k}{n})}$ for $n \in \mathbb{Z}$.[2] (if all eigenvalues are equally spaced on the unit circle) then the markov chain is non-absorbing and the power iteration does not converge to a fixed distribution as $t \to \infty$
5. The subspace generated by the power iteration is invariant[3] under P.
6. Each distribution over states x_t can be uniquely written as a sum of components, where each component lies in a subspace span$\{b_i, Pb_i, P^2b_i, \dots\}$ generated by applying P repeatedly to some base distributions b_i.

   Now consider the discounted markov chain

$$Q = I + \gamma P + \gamma^2 P^2 + ... + \gamma^n P^n + ..$$

[1] The power iteration is also called the krylov subspace.

[2] In an extreme periodic case, the eigenvalues can be evenly spaced on the unit circle (like roots of unity). In that case there is no single dominant direction, and the distribution cycles rather than converging.

[3] An invariant subspace S of P is a subspace such that for any vector u in S, Pu still lies in S. In other words, every distribution over states that lives in S stays in S under after one Markov step P.

Expanding this in terms of the eigendecomposition:

$$Q = V(I + \gamma\Lambda + \gamma^2\Lambda^2 + ...)V^{-1}$$

Each eigenvalue can be written as the series:

$$1 + \gamma\lambda_k + \gamma^2\lambda_k^2 + \cdots = \frac{1}{1 - \gamma\lambda_k},$$

And therefore the eigenvalues form the diagonal matrix:

$$\text{diag}\left(\frac{1}{1-\gamma\lambda_1}, \dots, \frac{1}{1-\gamma\lambda_n}\right).$$

Thus Q has the closed-form solution

$$Q = V\,\text{diag}\left(\frac{1}{1-\gamma\lambda_1}, \dots, \frac{1}{1-\gamma\lambda_n}\right)V^{-1} = (I - \gamma P)^{-1}.$$

P^k gives, for each starting state, the distribution over states after k steps (where you could be and with what probability). The matrix $Q = (I - \gamma P)^{-1}$ encodes the discounted sum of these k-step distributions. In other words, as we look k steps into the future, each step's contribution is downweighted by γ^k and then all of them are added together.[3]

We take the expectation over the discounted visits to a state since it is a random process and every start point $s_0 = i$ produces different rows for the Q matrix. We will let the matrix M be the summary of how that random process behaves from each starting point s_0:

$$M_{ij} = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \mathbf{1}\{s_t = j\} | s_0 = i)\right)$$

where $\mathbf{1}\{s_t = j\}$ is the indicator function:

$$\mathbf{1}\{s_t = j\} = \begin{cases} 1 & s_t = j \\ 0 & \text{otherwise.} \end{cases}$$

M represents an idealized summary of how discounted state occupancies evolve over time. Each row describes, for a given starting state, how often (discounted by γ^t) the process is expected to visit every other state. Since we do not have access to this M, we must learn an approximation to it, $\hat{M}$, through experience.

Once we have $\hat{M}$, we can compute some value v that relies on feedback from the environment about a reward.

$$v = \hat{M}\bar{r}.$$

[3] Each entry can also be represented in terms of conditional probabilities: $P_{ij} = Pr(S_{t+1} = j \mid S_t = i)$.

where $\bar{r}$ is the instantaneous feedback vector of expected immediate rewards at each state. If $\bar{r}$ is a basis vector this means at the state corresponding to the entry 1 yields a reward. If $\bar{r}$ has nonzero entries across many states, the value function mixes information about visits to those states.

For learning to make sense, we need the Markov chain to converge instead of wandering chaotically forever. If we iterate the chain long enough the distribution over states stabilizes to a steady pattern of how often each state is visited. The successor representation M captures a summary of these future visits. Since we down-weight time steps by γ^t, the infinite sum stays finite and behaves like a stable, long-term "footprint" of where the system tends to go from each starting state.

To capture this stability we would therefore want to minimize

$$\|m(s_t) - (e_{s_t} + \gamma \cdot m(s_{t+1}))\| \to 0$$

where e_{s_t} is a basis vector[4].

Proof: Each vector of $m(s_t) = (I - \gamma P)^{-1} e_{s_t}$ and $m(s_{t+1}) = (I - \gamma P)^{-1} e_{s_{t+1}}$. (I can finish this proof via bellman equations later...).

The learning rule is:

$$m(s_t) \leftarrow m(s_t) + \alpha \left[e_{s_t} + \gamma \cdot m(s_{t+1}) - m(s_t) \right]$$

Now to provide an update to the reward $\bar{r}$ at time t we will write

$$v(s_t) \approx \bar{r}_t + \gamma \cdot v(s_{t+1})$$

This means that our prediction at s_t is equal to the reward obtained at time t plus prediction at s_{t+1}.

$$\delta_t = \bar{r}_t + \gamma \cdot \underbrace{v(s_{t+1})}_{\text{target}} - \underbrace{v(s_t)}_{\text{current}}$$

to update $v(s_t)$ we will define:

$$v(s_t) \leftarrow v(s_t) + \alpha \cdot \delta_t$$

Full learning rule:

$$v(s_t) \leftarrow v(s_t) + \alpha \cdot (\bar{r}_t + \gamma \cdot v(s_{t+1}) - v(s_t))$$

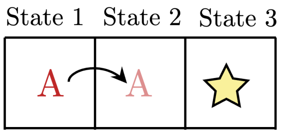
Concrete Example. We use the TD(0) update

$$v_{t+1}(s_t) = v_t(s_t) + \alpha\big(r_t + \gamma\, v_t(s_{t+1}) - v_t(s_t)\big),$$

with $\gamma = 0.9$, $\alpha = 0.5$, and initial values

$$v_0(1) = v_0(2) = v_0(3) = 0.$$

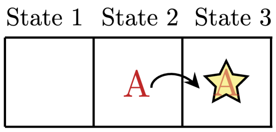
[4] We use a basis vector since this says we are at our current state s_t with probability one.



Step 1: transition 1 → 2 with reward $r_0 = 0$.

$$\begin{aligned}\delta_0 &= r_0 + \gamma\, v_0(2) - v_0(1) = 0 + 0.9 \cdot 0 - 0 = 0,\\ v_1(1) &= v_0(1) + \alpha\, \delta_0 = 0 + 0.5 \cdot 0 = 0.\end{aligned}$$

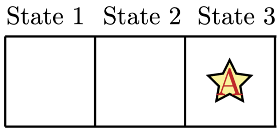
Other values unchanged: $v_1(2) = 0$, $v_1(3) = 0$.



Step 2: transition 2 → 3 with reward $r_1 = 0$.

$$\begin{aligned}\delta_1 &= r_1 + \gamma\, v_1(3) - v_1(2) = 0 + 0.9 \cdot 0 - 0 = 0,\\ v_2(2) &= v_1(2) + \alpha\, \delta_1 = 0 + 0.5 \cdot 0 = 0.\end{aligned}$$

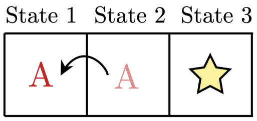
Again $v_2(1) = 0$, $v_2(3) = 0$.



Step 3: transition 3 → 3 with reward $r_2 = 1$.

$$\begin{aligned}\delta_2 &= r_2 + \gamma\, v_2(3) - v_2(3) = 1 + 0.9 \cdot 0 - 0 = 1,\\ v_3(3) &= v_2(3) + \alpha\, \delta_2 = 0 + 0.5 \cdot 1 = 0.5.\end{aligned}$$

So $v_3(1) = 0$, $v_3(2) = 0$, $v_3(3) = 0.5$.



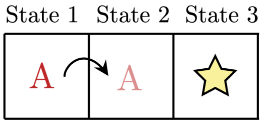
Step 4: transition $1 \to 2$ with reward $r_3 = 0$.

Note that after a reward, the episode ends and the transition starts from 1 again

$$\delta_3 = r_3 + \gamma\, v_3(2) - v_3(1) = 0 + 0.9 \cdot 0 - 0 = 0,$$

$$v_4(1) = v_3(1) + \alpha\, \delta_3 = 0.$$

Now $v_4(1) = 0$, $v_4(2) = 0$, $v_4(3) = 0.5$.

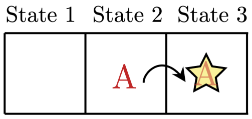


Step 5: transition $2 \to 3$ with reward $r_4 = 0$.

$$\delta_4 = r_4 + \gamma\, v_4(3) - v_4(2) = 0 + 0.9 \cdot 0.5 - 0 = 0.45,$$

$$v_5(2) = v_4(2) + \alpha\, \delta_4 = 0 + 0.5 \cdot 0.45 = 0.225.$$

So $v_5(1) = 0$, $v_5(2) = 0.225$, $v_5(3) = 0.5$.



Step 6: transition $3 \to 3$ with reward $r_5 = 1$.

$$\delta_5 = r_5 + \gamma\, v_5(3) - v_5(3) = 1 + 0.9 \cdot 0.5 - 0.5 = 1 + 0.45 - 0.5 = 0.95,$$

$$v_6(3) = v_5(3) + \alpha\, \delta_5 = 0.5 + 0.5 \cdot 0.95 = 0.975.$$

After these transitions,

$$v_6(1) = 0, \qquad v_6(2) = 0.225, \qquad v_6(3) = 0.975.$$