- an episode may never finish.

Then, it is possible that iterative policy evaluation will not converge, and this is because the state-value function may not be well-defined! To see this, note that in this case, calculating a state value could involve adding up an infinite number of (expected) rewards, where the sum may not converge.

In case it would help to see a concrete example, consider an MDP with:

- two states s_1 and s_2, where s_2 is a terminal state
- one action a (*Note: An MDP with only one action can also be referred to as a* ***Markov Reward Process (MRP)***.)
- $p(s_1, 1|s_1, a) = 1$

In this case, say the agent's policy π is to "select" the only action that's available, so $\pi(s_1) = a$. Say $\gamma = 1$. According to the one-step dynamics, if the agent starts in state s_1, it will stay in that state forever and never encounter the terminal state s_2.

In this case, $v_\pi(s_1)$ **is not well-defined**. To see this, remember that $v_\pi(s_1)$ is the (expected) return after visiting state s_1, and we have that

$$v_\pi(s_1) = 1 + 1 + 1 + 1 + \ldots$$

which diverges to infinity. (Take the time now to convince yourself that if either of the two convergence conditions were satisfied in this example, then $v_\pi(s_1)$ would be well-defined. As a **very optional** next step, if you want to verify this mathematically, you may find it useful to review geometric series and the negative binomial distribution.)