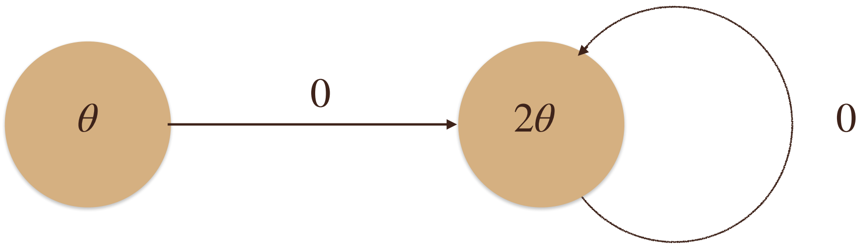

Figure 1. A counterexample shows that the fixed point of a TD model with fewer parameters than the number of states does not exist. In this example, each state has exactly one action and rewards are labelled next to the transitions. Value functions are parameterized by $\theta in \mathbb{R}$ as shown in the graph.

Why over-parameterized model is needed?

The feature matrix of these two states equals to $\Phi = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. The transition matrix is denoted by P. Let the diagonal of the matrix D be some state distribution. The learning is off-policy if the state distribution differs from the stationary distribution, which concentrates on the right state with self-loop. For any discount factor $\gamma > 0.5$, the state distribution $(\frac{4\gamma-4}{2\gamma-3}, \frac{1-2\gamma}{2\gamma-3})$ causes $\Phi^\top D(I - \gamma P)\Phi = 0$. Therefore, the fixed point $(\Phi^\top D(I - \gamma P)\Phi)^{-1}\Phi^\top DR$ does not exist. TD with or without a target network will be stuck at any initialization.