- **Gambler's-Ruin (Analytic Closed Form):** Because this is a symmetric random walk with absorbing states at the ends, the probability of eventually reaching the right terminal from state s is $\frac{s}{6}$, and thus

$$V(s) = 1 \cdot \left(1 - \tfrac{s}{6}\right) + 10 \cdot \left(\tfrac{s}{6}\right) = 1 + 9\,\frac{s}{6}.$$

- **Linear-System Solve (Policy Evaluation):** Under the fixed (random) policy, the Bellman equations for the five interior states can be written in matrix form as

$$(I - P)\,V = R,$$

  where P is the one-step transition matrix under the policy and R the expected immediate-reward vector. Solving this linear system yields the exact value function for the MRP.

Do you think the conclusions about which algorithm is better would be affected by a wide range of α values? Is there a different, fixed value of α at which either algorithm performs significantly well? Why or why not?

TD(0) outperforms constant-α First-Visit Monte Carlo holds across a wide range of learning rates. Though MC's bias–variance tradeoff could improve at very small values of α, its end-of-episode updates remain inherently high-variance and slow, making it never surpass TD(0). There is no single fixed α at which MC significantly outperforms TD(0), since TD(0)'s bootstrapping leads to lower variance and more resistance to the choice of α.

Produce another figure, where V(s) is drawn for each of the five non-terminal states individually. Now assume the parameter α decays from 0.5 over 250 episodes to 0.01. Compare TD(0) and the First Visit MC, what do you observe?

Comparing TD(0) to MC we can see that over time (as α_t gets smaller and smaller), TD(0) jumps less and less, while MC still peaks and falls a lot. At the 250th episode however, both methods seem to have estimated the V(s) fairly accurately, with sometimes MC being closer and sometimes TD(0) being closer. With a lower α_{tmax} and a higher t_{max} we would probably see one of the algorithms get closer to the true V(s) than the other.