	AA	CT	Elev	GOL	Nav	Recon	ST	Sysadm	Tam	Traffic	TT	Wildfire
Monte Carlo	-131.88	-22.27	-109.67	299.69	-37.27	0.0	8.07	614.09	-579.73	-49.22	40.78	-571.13
MaxMCTS(λ) at λ-value	**-127.78** 0.7	-21.49 0.4	***-83.95*** 0.0	299.69 1.0	-36.79 0.8	***1.67*** 0.3	8.07 1.0	614.09 1.0	-577.62 0.8	-48.71 0.9	**65.94** 0.7	-551.25 0.9
MaxMCTS$_\gamma$	**-128.53**	-22.3	-109.67	298.4	-36.87	**0.53**	6.16	610.43	-582.33	-48.77	**64.19**	-578.05
MCTS(λ) at λ-value	-131.69 1.0	-22.1 0.4	-109.67 0.0	299.97 1.0	-37.16 0.5	**0.04** 0.0	8.83 1.0	615.68 1.0	-573.07 0.1	-49.17 1.0	44.4 0.7	-515.8 0.6
MCTS$_\gamma$	-132.19	-22.5	-109.67	297.97	-37.3	0.0	9.29	614.33	-572.2	-49.16	40.11	-557.74

Table 2. Mean performance of different backup strategies for all 12 IPC domains. A value is bold if it is significantly better than Monte Carlo, i.e. MaxMCTS($\lambda = 1$). A value is italicized if it is significantly better than MaxMCTS$_\gamma$.

Next, we study the connection between domain structure and choice of backup strategy using a grid-world domain.

4.2. Grid World

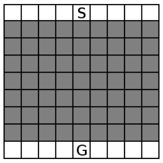
It is evident from the results across the IPC domains that the same complex backup strategy can lead to vastly different performance across particular domains. How is domain structure linked to different performance for different backup strategies, and can we use this structure to inform our choice of backup strategy? To shed more light on these questions, we perform multiple tests on a controlled grid-world environment, and obtain a deeper understanding of the MCTS mechanics given different conditions.

All of our tests are conducted on a grid-world environment of size 9x9 which has to be navigated by an agent, as depicted in Figure 4a. The agent's start and goal locations are fixed and labeled as S and G in the figure, respectively. The domain supports 4 non-deterministic actions: *up*, *down*, *left*, and *right* with the following transitions:

$$p(ns_a|a,s) = 0.925, \text{ and } p(ns_{a'}|a,s) = 0.025 \mid a' \neq a,$$

where $a, a' \in \{up, down, left, right\}$, s is the agent's current cell, ns_a is the adjacent cell in direction a. If the agent is next to an edge, the probability mass of moving into the edge is mapped into the current cell of the agent. When the agent reaches the goal state, it is given a reward of +100, and all other actions result in a reward of -1. The MCTS search depth parameter, *planHorizon*, is set to 100. Planning is done using 10,000 simulations with uniform action selection, and the previous search tree is reused.

Our main hypothesis regarding the performance of different backup strategies revolves around value estimation during exploration. More exactly, we believe it is a matter of how many optimal or close-to-optimal trajectories exist in the tree. In the grid-world environment, there are cases when there are a lot of paths in the search tree that the agent can use to get to the goal. Consequently, using a Monte Carlo backup produces the best results (Figure 4b, $\lambda = 1$ with no 0-reward terminal states), as other backup approaches have a greater likelihood of getting stuck in suboptimal trajectories due to an off-policy update.



(a) Grid

#0-Term	0	3	6	9	12	15	18
$\lambda = 1$	***90.4***	11.3	0.9	-1.3	-2.1	-2.2	-2.2
$\lambda = 0.8$	90.2	28.0	10.7	5.9	0.3	-1.4	-2.0
$\lambda = 0.6$	89.5	62.8	45.3	30.6	17.5	8.5	3.3
$\lambda = 0.4$	88.7	***85.1***	77.6	62.2	41.7	24.1	10.1
$\lambda = 0.2$	87.7	82.6	**78.1**	***69.9***	**51.3**	28.4	13.2
$\lambda = 0$	84.5	79.8	74.1	67.0	50.2	**31.8**	**15.75**
γ	90.1	25.7	15.3	13.2	8.2	4.7	2.3

(b) Varying 0-Reward Terminal States

Figure 4. The gray cells in (a) indicate the cells that can become 0-reward terminal states. (b) tabulates average episodic reward as the number of 0-reward terminal states is varied. The highest performing backup strategy for each domain setting is bold, and also italicized if significantly better than all other column entries.

On the other hand, if there are very few successful paths to the goal, then averaging out multiple rollouts can easily drown value backups from rare close-to-optimal trajectories. We test this hypothesis by introducing a number of additional terminal states, or barriers, between the start and goal locations in the grid formulation. These barriers are chosen at random from the grey cells in Figure 4a. If the agent transitions to one of these cells, it receives a reward of 0, and the episode terminates. The average episodic reward obtained by MaxMCTS(λ) and MaxMCTS$_\gamma$ as the number of terminal states varied is tabulated in Figure 4b.

When the number of obstacles is 0, Monte Carlo or MaxMCTS(1) significantly outperforms all other approaches. However, with only three obstacles, the performance of MaxMCTS$_\gamma$ and MaxMCTS($\lambda \geq 0.6$) deteriorates sharply, and MaxMCTS(0.4) significantly outperforms all other approaches. The deterioration in performance for