Appendix 1.A Detailed Experimental Results for the SensorPlacement Problems

Table 2. Average total discounted rewards and 95% confidence intervals of all tested solvers on the SensorPlacement problems. The average is taken over 1000 simulation runs per solver and problem, with a planning time of 1s per step.

	SensorPlacement-6	SensorPlacement-8	SensorPlacement-10	SensorPlacement-12
ADVT	**842.8 ± 9.5**	**706.8 ± 17.5**	**565.1 ± 21.7**	**303.0 ± 19.8**
ADVT-R	676.3 ± 19.7	238.1 ± 33.5	28.7 ± 18.4	-17.3 ± 7.4
ADVT ($L = 0$)	780.4 ± 12.6	448.8 ± 15.9	325.3 ± 16.2	102.5 ± 7.4
ADVT-MC	812.6 ± 11.4	692.7 ± 17.7	551.2 ± 18.3	293.5 ± 19.6
VOMCPOW-B	823.4 ± 15.1	679.1 ± 17.9	481.6 ± 22.2	191.3 ± 17.6
VOMCPOW-I	817.2 ± 15.8	663.4 ± 18.6	476.0 ± 22.7	189.9 ± 18.0
VOMCPOW	768.5 ± 16.4	305.6 ± 25.8	110.1 ± 24.5	-8.2 ± 13.2
POMCPOW-B	659.3 ± 17.2	428.7 ± 21.5	114.6 ± 16.6	-1.9 ± 6.6
POMCPOW-I	653.2 ± 17.3	425.2 ± 21.8	111.3 ± 16.8	-2.1 ± 6.8
POMCPOW	377.6 ± 23.5	113.4 ± 24.2	-36.8 ± 11.3	-74.3 ± 12.9

Appendix 1.B Success Rates

Table 3. Success rates of all tested solvers on the Pushbox, Parking and VDP-Tag problems. The success rate is with respect to 1,000 simulation per solver and problem, with a planning time of 1s per step.

	Pushbox2D	Pushbox3D	Parking2D	Parking3D	VDP-Tag
ADVT	0.985	0.969	0.912	0.916	0.941
ADVT-R	0.987	0.968	0.943	0.906	0.945
ADVT (L=0)	0.966	0.965	0.857	0.898	0.935
ADVT-MC	0.989	0.972	0.417	0.337	0.892
VOMCPOW-B	0.985	0.970	0.885	0.886	-
VOMCPOW-I	0.975	0.939	0.597	0.314	-
VOMCPOW	0.754	0.815	0.512	0.297	0.987
POMCPOW-B	0.974	0.953	0.853	0.534	-
POMCPOW-I	0.963	0.969	0.409	0.122	-
POMCPOW	0.712	0.692	0.401	0.125	0.979

In addition to the average total discounted rewards in the main document, we also report the success rate of each solver in each problem scenario in Table 3 and Table 4. For the Pushbox problems, a run is considered successful if the robot pushes the opponent into the goal region, while avoiding collisions of itself and the opponent with the boundary region. For the Parking problems, a run is successful if the vehicle reaches the goal area. For the VDP-Tag problem a run is successful if the opponent is being tagged and for the SensorPlacement problems, a run is successful if the end-effector reaches the sensor mounting location. In all problems, the task must be completed within planning steps 50 steps, otherwise problem terminates and the run is considered unsuccessful.

Generally the success rates are closely correlated to the average total discounted rewards achieved by each solver in the problem scenarios. The results in Table 4 further indicate that ADVT scales better to higher-dimensional action spaces compared to the baselines. However, for the SensorPlacement12 problem,

Table 4. Success rates of all tested solvers on the SensorPlacement problems. The success rate is with respect to 1,000 simulation per solver and problem, with a planning time of 1s per step.

	SensorPlacement-6	SensorPlacement-8	SensorPlacement-10	SensorPlacement-12
ADVT	0.981	0.962	0.834	0.724
ADVT-R	0.832	0.692	0.756	0.557
ADVT ($L = 0$)	0.937	0.726	0.791	0.601
ADVT-MC	0.964	0.959	0.828	0.719
VOMCPOW-B	0.979	0.951	0.807	0.703
VOMCPOW-I	0.967	0.891	0.803	0.698
VOMCPOW	0.923	0.721	0.645	0.583
POMCPOW-B	0.829	0.794	0.646	0.575
POMCPOW-I	0.826	0.781	0.657	0.578
POMCPOW	0.738	0.636	0.519	0.321

the success rates are relatively low, even for ADVT. Thus, for such problems, developing methods that scale even better to high-dimensional action spaces remains a fruitful avenue for future research.

Appendix 1.C Observation Discretization Method for the VDP-Tag Problem

Since the observation space in the VDP-Tag problem is continuous, i.e., $\mathcal{O} = \mathbb{R}^8$, ADVT requires a method to discretize the observations. To this end, we use a simple distance-based discretization: Suppose a sampled episode selects action $a \in \mathcal{A}(b)$ at belief b and perceives an observation $o_i \in \mathcal{O}$. We then compute the Euclidean distance between o_i and each observation corresponding to the outgoing edges $(a, o) \in \mathcal{T}$ that descend b. If there is an observation o_k corresponding to an outgoing edge for which the Euclidean distance to o_i yields a value smaller than a threshold δ (in our experiments we use $\delta = 25$), we continue the search from child node b' of b via edge (a, o_k). Otherwise, we add a new child node to b via edge (a, o_i).

Appendix 1.D Results for the VDP-Tag Problem With Smaller Transition Errors

Table 5. Average total discounted rewards with 95% confidence intervals and success rates of all tested solvers on the VDP-Tag with smaller transition errors ($\sigma = 0.01$). The average is taken over 1000 simulation runs per solver and problem, with a planning time of 1s per step.

	Avg. total discounted reward	Success rate
ADVT	31.5 ± 0.9	0.991
ADVT-R	31.7 ± 0.9	0.986
ADVT L(=0)	30.1 ± 1.0	0.989
ADVT-MC	30.9 ± 0.8	0.984
VOMCPOW	$\mathbf{34.1 \pm 0.8}$	**0.998**
POMCPOW	29.1 ± 0.8	0.990

We additionally tested ADVT as well as VOMCPOW and POMCPOW on a variant of the VDP-Tag problem with smaller transition errors, i.e., the position of the target is disturbed by Gaussian noise with standard deviation $\sigma = 0.01$ instead of $\sigma = 0.05$. The results are shown in Table 5.

It can be seen that for the variant of the problem with $\sigma = 0.01$, ADVT is competitive with POMCPOW. For $\sigma = 0.05$ the uncertainty with respect to the position of the target is large and therefore the agent must carefully decide when to activate its range sensor in order to reduce uncertainty. To achieve this, the solvers require more accurate belief representations in the search tree, which is challenging for ADVT, as it relies on discretizing the continuous observation space. On the other hand, VOMCPOW and POMCPOW use Progressive Widening in the observation space, combined with a weighted-particle representation of the beliefs in the search trees, which helps them to perform well.

For $\sigma = 0.01$, the uncertainty with respect to the position of the target is much smaller, and therefore the solvers are less reliant on accurate belief representations in the search tree, which benefits ADVT. This suggests that our method works well for problems with small belief uncertainties, even if the observation space is continuous. To handle larger uncertainties, we require a better method to handle continuous observation spaces, such as progressive widening and explicit belief representations as used in VOMCPOW and POMCPOW. We are planning to explore this in future works.