Table 2: The Rewards Gained by SARSOP and HSVI2 for Different Times on RockSample[7,8] and Homecare Compared to Pairwise Heuristic

Method	Average reward	Total time(s)
RockSample[7,8]		
Pairwise Heuristic	$18.76 \pm .23$	$0 + 0.03$
SARSOP	$7.35 \pm .00$	$0.22 + 0$
SARSOP	$17.57 \pm .30$	$0.37 + 0$
HSVI2	$10.43 \pm .00$	$2.00 + 0$
HSVI2	13.87 ± 0.12	$4.00 + 0$
Homecare		
Pairwise Heuristic	$15.79 \pm .81$	$0 + 0.45$
SARSOP	$14.44 \pm .67$	$230.12 + 0$
HSVI2	$14.36 \pm .52$	$250.00 + 0$

because we put zero for our method. Actually, the offline computation in HSVI2 and SARSOP is different from our offline computation. Unlike our method, the offline computation in HSVI2 and SARSOP is dependent on the initial belief state. As a result, they should be performed again each time the initial belief state changes. But, our offline part should be performed only once for each problem, no matter what the initial belief state is. Also, note that the total online time is the time needed to find and execute the entire plan; it is not the time required for just one step.

Table 1 shows that the pairwise heuristic is definitely a better approach for solving the Fourth problem and its time needed for solving RockSample[7,8] and Homecare is much less than the time needed for SARSOP and HSVI2. Furthermore, our method always gains much more reward than QMDP, entropy-weighting and also POMCP with the time limit of one minute. In regards to RockSample[7,8] and Homecare, one may argue that even though SARSOP and HSVI2 need more time to gain the optimum reward, they might gain the same or even higher reward than our method in its required time ($0.03s$ for Rocksample and 0.6 for Homecare). For Rocksample problem, we tested these methods in $0.22s$, 7 times our time needed and $0.37s$, 10 times our time needed for SARSOP and $2.00s$ and $4.00s$ for HSVI2 and show the results in Table 2. As shown, they achieve less reward in 10 times the time required by the pairwise heuristic.

For Homecare, SARSOP needs $208s$ only for initialization, showing that it cannot generate any policy in less than 460 times the time required by the pairwise heuristic. Initialization time is $244s$ for HSVI2. We report the reward of SARSOP in $230.12s$, and HSVI in $250.00s$ in Table 2 showing that these methods gain less reward in more than 500 times the time required by our method.

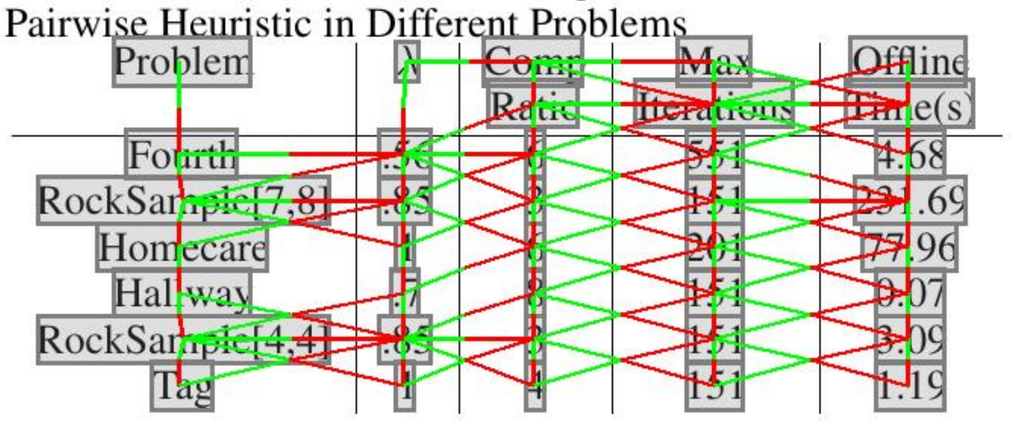
The reward and the time needed for the other problems are illustrated in Table 3. This table shows that the pairwise heuristic gains a near-optimal reward in all tested problems.

Table 4 shows the parameters and required offline time of the pairwise heuristic in all problems. The reported offline time shows that in some large problems even the sum of offline and online time required for the pairwise heuristic is less than the time SARSOP and HSVI2 require for solving one trial.

Table 3: The Average Reward and the Time Required for the Methods on Classical Small Problems

Method	Average reward	Offline time(s)	Online time(s)
Hallway			
Pairwise Heuristic	$.81 \pm .02$	0	0.01
QMDP	$.33 \pm .03$	0	0.01
EW	.60	0	NA
HSVI2	$1.01 \pm .03$	1.00	0
SARSOP	$.99 \pm .04$	0.30	0
RockSample[4,4]			
Pairwise Heuristic	$16.21 \pm .32$	0	0.01
POMCP	$14.15 \pm .31$	0	10.12
QMDP	$3.29 \pm .36$	0	0.02
HSVI2	$17.91 \pm .12$	1.00	0
SARSOP	$18.03 \pm .06$	0.56	0
Tag			
Pairwise Heuristic	-7.18 ± 0.25	0	0.03
POMCP	$-6.44 \pm .45$	0	12.74
QMDP	-16.55 ± 0.32	0	0.05
HSVI2	$-6.30 \pm .10$	7.00	0
SARSOP	$-6.12 \pm .15$	1.07	0

Table 4: The Parameters and required offline time of the Pairwise Heuristic in Different Problems



Analysis and Discussion

We tested the pairwise heuristics on classical test benchmarks in the POMDP literature and got near-optimal solutions in all of them. However, our approach does not always work well especially if reducing uncertainty is not essential for getting the maximum reward. In fact, the biggest drawback of the pairwise heuristic is that there is no lower bound for the reward of its solution. However, there is a bound for the removing uncertainty phase in some cases. Golovin et al. used a similar pairwise heuristic named EC^2 (Equivalence Class Edge Cutting) in the active learning field and got a near-optimal policy with the following bound:

$$C(\pi_{EC^2}) \leq (2\ln(1/p_{min}) + 1)C(\pi^*) \qquad (10)$$

π^* is the optimal policy and π_{EC^2} is the policy generated by EC^2 method. $C(\pi)$ is the total cost of a policy and