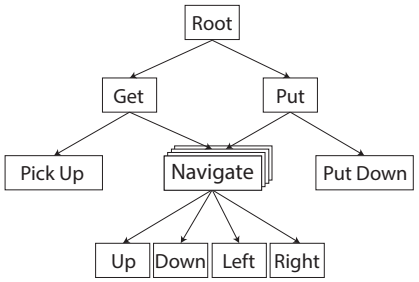

Figure 1: Task hierarchy for the taxi problem. There are four navigation subtasks, which can be represented using a single fragment.

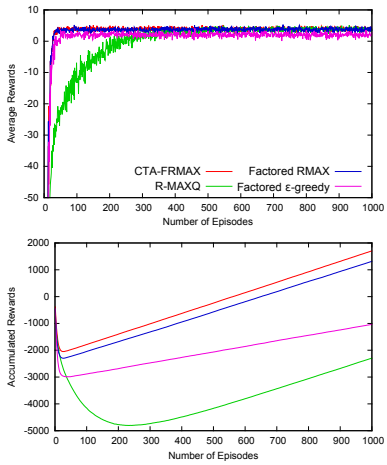

Figure 2: Results on MAXQ taxi.

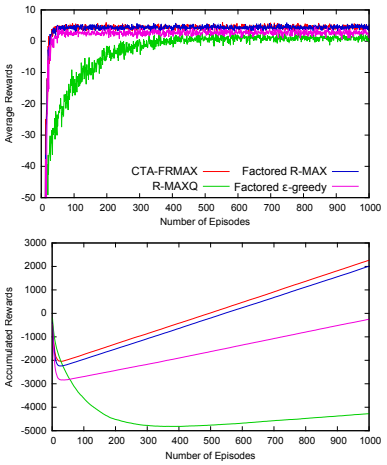

Figure 3: Results on fickle taxi.

model for node i is given by,

$$P(F_i'|F_i,a) = \prod_{Y:Y\subseteq N\wedge i\in Y} P\left(Uni(\bigcap_{j\in Y} F_j') \mid \bigcap_{j\in X_a\cap Y} F_j, a\right)$$

The theorem states that the transition function is a product of factors where, the parent set of $Uni(\bigcap_{j\in Y} F_j')$ depends on the action being executed. This is captured in the relation $\bigcap_{j\in X_a\cap Y} F_j$.

The direct outcome of the theorem is a reduced feature set for learning. We combine the safe state abstraction information and contextual independence to eliminate the maximum number of features from the parent set to improve the learning efficiency.

We introduce the CTA-FRMAX algorithm by combining efficient transition function learning in Theorem 1 with RMAX style exploration in factored spaces [3]. CTA-FRMAX maintains an exploration count on each factor in Theorem 1. If the count is smaller than a threshold, the agent transits to a fictitious state and receives maximum reward.

3. RESULTS

We test the performance of CTA-FRMAX in the original MAXQ taxi and the fickle taxi benchmark problems [2]. The experimental setup involves a 5×5 grid-world with four landmarks. The passenger location and destination must be one of the landmarks. There are six actions: *navigate* along the four directions, *pickup* and *putdown* as shown in Figure 1. In the fickle taxi variant the passenger may change the destination with probability 0.3. The episode ends if passenger reaches the destination or 1000 time steps elapse.

CTA-FRMAX computes the transition function for both *Get* and *Put*, and executes *Get* when the passenger is not on board and *Put* otherwise. Navigation is modeled as a fragment. We compare the performance of CTA-FRMAX against R-MAXQ [4], Factored RMAX [3] and Factored ϵ-greedy. The latter two methods are given the full dynamic Bayesian network (DBN) representation of the transition functions for both the *Get* and *Put*. We set the exploration threshold m to be 1 for both CTA-FRMAX and Factored RMAX. Since R-MAXQ does not converge with $m = 1$, we use $m = 5$ like other methods [4, 1].

As shown in Figure 2 and Figure 3, CTA-FRMAX outperforms all the other methods. In terms of number of episodes, CTA-FRMAX converges in about 50 episodes whereas R-MAXQ takes about 500 episodes to converge due to the delayed exploration—it is unable to explore *putdown* unless it completes the current navigation task. Also, R-MAXQ fails to converge to the optimal policy in the fickle taxi experiment, since it treats each subtask like a primitive action; it cannot immediately react to a change of destination until its current subtask completes. CTA-FRMAX also has better cumulative reward than Factored RMAX, despite that it only requires the relevant features for each task to be specified, while Factored RMAX requires the complete DBN structure of task transition dynamics.

Acknowledgments

This research is supported by a SMART Grant: Enhancing the Care for Elderly through Robotic Aides and Academic Research Grant: T1 251RES1211 from the Ministry of Education in Singapore.

REFERENCES

[1] F. Cao and S. Ray. Bayesian Hierarchical Reinforcement Learning. In *NIPS-12*, pages 73–81, 2012.

[2] T. G. Dietterich. The MAXQ Method for Hierarchical Reinforcement Learning. In *ICML-98*, pages 118–126, 1998.

[3] C. Guestrin, R. Patrascu, and D. Schuurmans. Algorithm-directed Exploration for Model-based Reinforcement Learning in Factored MDPs. In *ICML-02*, pages 235–242, 2002.

[4] N. K. Jong and P. Stone. Hierarchical Model-based Reinforcement Learning: R-MAX + MAXQ. In *ICML-08*, pages 432–439. ACM, 2008.

[5] R. Sutton, D. Precup, and S. Singh. Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112:181–211, 1999.

[6] N. A. Vien and M. Toussaint. Hierarchical Monte-Carlo Planning. In *AAAI-15*, 2015.