High-Quality Policies for the Canadian Traveler's Problem (Extended Abstract)

Patrick Eyerich and **Thomas Keller** and **Malte Helmert**
Albert-Ludwigs-Universität Freiburg
Institut für Informatik
Georges-Köhler-Allee 52
79110 Freiburg, Germany
{eyerich,tkeller,helmert}@informatik.uni-freiburg.de

The Canadian Traveler's Problem

The *Canadian Traveler's Problem* (CTP; Papadimitriou and Yannakakis 1991) is a path planning problem with imperfect information about the roadmap. We consider its *stochastic* version, which has drawn considerable attention from researchers in AI search (e. g., Nikolova and Karger 2008) and is closely related to navigation tasks in uncertain terrain considered in the robotics literature (e. g., Koenig and Likhachev 2002; Likhachev and Stentz 2006).[1]

The objective in the CTP is to travel from the initial location v_0 to some goal location $v_\star$ on a roadmap given as an undirected weighted graph with vertex set V (*locations*) and edge set E (*roads*). Complicating things, only a subset of roads $W \subseteq E$, called the *weather*, is actually traversable. The weather remains static while the agent traverses the graph. The agent does not know the weather; however, it does knows with which probability each road is blocked, and it automatically observes the status of a road upon reaching one of the incident locations. We speak of *good* weather when v_0 and $v_\star$ are connected under weather W and *bad* weather otherwise. Instances of the stochastic CTP are examples of *deterministic POMDPs* (Bonet 2009), where partial observability of the initial state (due to unknown weather) is the only source of uncertainty.

Policies for the CTP

We describe and evaluate four algorithms for computing policies for the stochastic CTP. More precisely, we describe four algorithms for computing *cost functions* for belief states of the stochastic CTP. Each of these then induces a policy which acts greedily with respect to the respective cost function. The first algorithm we consider, which dominates the literature on the stochastic CTP, is the *optimistic* policy (OMT), which is based on what is called the *free space assumption* in the robotics literature: as long as it is *possible* that a given road is traversable, we assume that it is traversable. Formally, the optimistic cost function in belief state b, $C_{\text{OMT}}(b)$, is the distance from the agent location to the goal in the *optimistic roadmap* for b, which is the graph that includes all roads that are not known to be impassable in b. Finding shortest paths in the optimistic roadmap is a standard shortest path problem without uncertainty, and hence $C_{\text{OMT}}(b)$ can be efficiently computed.

Our second approach, which still allows us to reduce cost estimation to shortest path computation in regular graphs is *hindsight optimization* (HOP). For each belief state, N iterations called *rollout* are performed, where N is a parameter. In each rollout, a weather W that is consistent with the current belief state b is sampled according to the blocking probabilities. If W is bad, the rollout counts as failed. Otherwise, we compute the goal distance of the agent location in the graph $\langle V, W \rangle$. The HOP cost estimate $C^N_{\text{HOP}}(b)$ for N rollouts is the average of the computed distances over all successful rollouts. An alternative name for HOP is *averaging over clairvoyance*, since its rollouts assume a "clairvoyant" agent that knows ahead of time which roads are traversable.

This assumption of clairvoyance is the Achilles heel of HOP. Our third approach, *optimistic rollout* (ORO) addresses this issue by modifying how each rollout is performed. While its cost function C^N_{ORO} is also computed by performing a sequence of N rollouts and averaging over cost estimates for successful rollouts, rather than using the clairvoyant goal distance in good weather W, ORO *simulates the optimistic policy* on W and uses the cost of the resulting run as the cost estimate.

The final approach we consider is the *UCT* algorithm (Kocsis and Szepesvári 2006). As in ORO, each UCT rollout computes an actual run from the agent location to the goal for the given weather, without using information that is hidden to the agent, and uses the average cost of successful rollouts as the overall cost estimate $C^N_{\text{UCT}}(b)$. While each rollout is independent in HOP and ORO, UCT rollouts are correlated: when deciding which successor to pick, UCT favors those that led to *low cost* and have been *rarely tried* in previous rollouts. To balance these criteria, i. e., the classical trade-off between exploitation and exploration, it picks a candidate maximizing the *UCT formula*, which is designed to select each successor arbitrarily often in the limit while choosing previously unpromising successors increasingly more and more rarely over time.

The basic or *blind* UCT algorithm (UCTB) does not take any problem-specific information into account to bias rollouts towards the goal and performs poorly in our experi-

[1] The work described in this extended abstract is presented in detail in a AAAI 2010 paper (Eyerich, Keller, and Helmert 2010).

	OMT	HOP	ORO	UCTB	UCTO
$\varnothing\, C_{20}$	186.5±2	165.7±2	162.2±2	185.9±2	**155.4**±2
$\varnothing\, C_{50}$	303.7±3	275.4±3	259.5±3	392.1±6	**244.7**±2
$\varnothing\, C_{100}$	383.0±5	371.3±4	331.5±4	564.9±10	**316.5**±3

Table 1: Average travel costs with 95% confidence intervals for 1000 runs on each of 10 roadmaps with 20, 50, and 100 locations. Best results are highlighted in bold.

ments. Hence, we also consider the *optimistic UCT* (UCTO) approach, which modifies UCTB in two ways:

- When extending a partial rollout which has several unvisited successors, pick the one with lowest C_{OMT} value.
- When evaluating the UCT formula, pretend that there were M additional rollouts for each successor, each with cost $C_{\text{OMT}}(b)$, where M is an algorithm parameter.

These modifications guide early rollouts towards promising parts of the belief space while not affecting the behavior in the limit. Similar extensions to UCT have shown great success in the game of Go (Gelly and Silver 2007).

Evaluation

We evaluated the presented CTP algorithms theoretically and experimentally. A first, fairly direct theoretic results is that the cost functions of all considered randomized algorithms (HOP, ORO and both UCT variants) converge in probability, so that it makes sense to speak of the cost functions "in the limit", which we denote as C^∞_{HOP}, C^∞_{ORO} and C^∞_{UCT}. (There is no need to distinguish UCTB and UCTO here, since their behavior in the limit is identical.) Using these notations, we can present our main result:

Theorem 1 *For all CTP instances $\mathcal{I}$ and belief states b:*

$$C_{OMT}(b) \leq C^\infty_{HOP}(b) \leq C^\infty_{UCT}(b) = C^*(b) \leq C^\infty_{ORO}(b),$$

where $C^*(b)$ is the cost function of the optimal policy. Moreover, there are instances where all inequalities are strict and the ratio between any two different cost functions is arbitrarily large.

The theorem implies that UCT converges towards an optimal policy. It also confirms the intuition that HOP (and even more so OMT) is an optimistic approximation to the real cost function while ORO is pessimistic.

To test how theory matches actual performance, we conducted an experiment with 30 problem instances based on random Delaunay graphs with 20, 50 or 100 locations and up to 287 roads. We estimate the expected cost of a policy by performing 1000 runs for each algorithm and benchmark graph, using $N = 10000$ rollouts for all randomized algorithms and $M = 20$ for UCTO. Table 1 shows that UCTO dominates, always providing the cheapest policies, but HOP and ORO also significantly outperform OMT, clearly demonstrating the benefit of taking uncertainty into account. On the middle-sized graphs, UCTO improves over OMT by 19.4% on average, with similar results on the remaining benchmarks. UCTB, on the other hand, does not fare well, converging too slowly and thereby underlining that these benchmarks are far from trivial. This result is not unexpected, as the initial rollouts of UCTB have to reach the goal through random walks.

To analyze the speed of convergence and scalability, we performed additional experiments on some of the benchmark graphs where we varied the rollout number in the range 10–100000. After only about 100 rollouts, all randomized algorithms except UCTB already obtain a better quality than OMT. After about 1000 rollouts, ORO and HOP begin to level off while UCTO takes the lead and still continues to improve. The experiment also shows that the eventual convergence of UCTB to an optimal policy is of limited practical utility as the speed of convergence is very low.

In order to also provide a comparison to related approaches suggested in the literature, we did an additional experiment on the *CTP with remote sensing*, where the agent has the additional option of sense the status of roads from a distance at a cost. (For the regular stochastic CTP, we found no other work to directly compare to.) We applied our algorithms to this setting by never making use of remote sensing. The resulting policies are competitive to sensing policies suggested in the literature (Bnaya, Felner, and Shimony 2009) and even outperform them on instances with larger blocking probabilities.

Conclusion

We investigated the problem of finding high-quality policies for the stochastic CTP. We discussed several algorithms for this problem that outperform the optimistic approach that is prevalent in the literature, both in terms of theoretical guarantees and in terms of experimental solution quality. The best-performing algorithm, UCTO, converges to an optimal policy in the limit and offers very good empirical performance with a reasonable number of rollouts.

References

Bnaya, Z.; Felner, A.; and Shimony, S. E. 2009. Canadian traveler problem with remote sensing. In *Proc. IJCAI 2009*, 437–442.

Bonet, B. 2009. Deterministic POMDPs revisited. In *Proc. UAI 2009*, 59–66.

Eyerich, P.; Keller, T.; and Helmert, M. 2010. High-quality policies for the Canadian travelers problem. In *Proc. AAAI 2010*, 51–58.

Gelly, S., and Silver, D. 2007. Combining online and offline knowledge in UCT. In *Proc. ICML 2007*, 273–280.

Kocsis, L., and Szepesvári, C. 2006. Bandit based Monte-Carlo planning. In *Proc. ECML 2006*, 282–293.

Koenig, S., and Likhachev, M. 2002. D* Lite. In *Proc. AAAI 2002*, 476–483.

Likhachev, M., and Stentz, A. 2006. PPCP: Efficient probabilistic planning with clear preferences in partially-known environments. In *Proc. AAAI 2006*, 860–867.

Nikolova, E., and Karger, D. R. 2008. Route planning under uncertainty: The Canadian traveller problem. In *Proc. AAAI 2008*, 969–974.

Papadimitriou, C. H., and Yannakakis, M. 1991. Shortest paths without a map. *Theoretical Computer Science* 84(1):127–150.