Toward a Topological Theory of Relational Reinforcement Learning for Navigation Tasks

Terran Lane
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
terran@cs.unm.edu

Andrew Wilson
Sandia National Laboratories
Albuquerque, NM 87131
atwilso@sandia.gov

Abstract

We examine application of relational learning methods to reinforcement learning in spatial navigation tasks. Specifically, we consider a goal-seeking agent with noisy control actions embedded in an environment with strong topological structure. While formally a Markov decision process (MDP), this task possesses special structure derived from the underlying topology that can be exploited to speed learning. We describe relational policies for such environments that are relocatable by virtue of being parameterized solely in terms of the relations (distance and direction) between the agent's current state and the goal state. We demonstrate that this formulation yields significant learning improvements in completely homogeneous environments for which exact policy relocation is possible. We also examine the effects of non-homogeneities such as walls or obstacles and show that their effects can be neglected if they fall outside of a closed-form envelope surrounding the optimal path between the agent and the goal. To our knowledge, this is the first closed-form result for the structure of an envelope in an MDP. We demonstrate that relational reinforcement learning in an environment that obeys the envelope constraints also yields substantial learning performance improvements.

Introduction

While the field of reinforcement learning (RL) has achieved a number of impressive successes, RL methods still suffer from slow convergence in many domains and have not yet found general, widespread acceptance in many apparently ideal RL domains. Recent years have seen the advent of relational methods across a wide spectrum of learning tasks including data mining (Getoor *et al.* 2001), web navigation analysis (Anderson, Domingos, & Weld 2002), and reinforcement learning (Boutilier, Dearden, & Goldszmidt 2000; Finney *et al.* 2002).

In this paper, we argue that navigational tasks in geographic environments, e.g., the kinds of tasks encountered by robots moving through the physical world, possess special structure that renders them particularly well suited to relational methods. In particular, large subsets of the physical world are characterized by *locality*, *homogeneity*, and *translation/rotation invariances*. Furthermore, unlike complex combinatorial planning problems such as blocksworld (Finney *et al.* 2002) or task planning (Boutilier, Dearden, & Goldszmidt 2000), navigational domains are typically not prone to exponential explosions of trajectories. These properties have been well understood and exploited for hundreds of years in the context of deterministic motion planning, but it is much less clear how to interpret such properties in the kinds of stochastic domains that we typically deal with in RL. The fundamental difficulty is that our understanding of deterministic motion planning is rooted in the metric of physical spaces — the important properties of a space depend only on the *distance* between points and absolute coordinates are irrelevant — while RL methods are typically based on the Markov decision process (MDP) formalism in which policies are tied to atomic states, i.e., to absolute coordinates.

In this paper we take first steps toward developing a relational theory for RL in navigational Markov decision processes. We wish to be able to describe navigational tasks in terms of something like a distance or orientation between states. That is, "for all states located X units south and Y units west of the goal state, act as follows..." We exploit recent results by Ravindran and Barto (Ravindran & Barto 2002; 2003; Ravindran 2004) on the equivalence of options (partial policies) under homeomorphic transforms to construct goal-seeking policies described only in terms of the relationship between current state and goal state, independently of absolute coordinate. The difficulty is that, even when the underlying space is metric and homogeneous, the behavior of a stochastic agent may *not* be exactly relocatable. Obstacles, even off of the "optimal trajectory" between current and goal state, may distort transition probabilities and, therefore, typical notions of distance.

In response to this problem, we develop a closed-form bound on the high-probability envelope (Dean *et al.* 1995) of trajectories that an agent can take in traversing between two states while executing a fixed policy. This bound is described in terms of the topology of the underlying space, allowing us to *a priori* describe the states that an agent may reasonably enter en route to the goal state. To our knowledge, this is the first closed-form description of a policy envelope for an MDP. Such an envelope allows us to describe when a goal-seeking policy is "approximately relocatable". Essentially, when no obstructions fall within the envelope, the probability that the agent's trajectories (and, therefore,

its value function) will be distorted by obstructions, is low enough to neglect. Unlike methods that have been developed for MDPs with "rooms" or "corridors" (Hauskrecht *et al.* 1998; Parr 1998a), our results are most applicable to open and largely unobstructed spaces such as free space or open field navigation.

This paper does not aim to be a *complete* or *general* solution to navigational problems in relational navigation domains. Rather, it is a first step toward such a theory. Our primary concerns here are developing a useful relational representation for navigational MDPs, describing important topological properties of such MDPs, and demonstrating that those properties lead to strong constraints on an agent's behaviors.

Background and Definitions

We give here only a very brief review of the notation and key aspects of Markov decision processes and reinforcement learning in them. For a thorough treatment, we refer the reader to the texts by Puterman (Puterman 1994) and Sutton and Barto (Sutton & Barto 1998), respectively.

An MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, R \rangle$ is a stochastic control process specified by four components: a *state space*, $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, of cardinality $|\mathcal{S}| = N$ (taken to be discrete and finite in this paper); a set of primitive (or atomic) *actions*, $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ (also finite); a *transition function*, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$; and a *reward function*, $R : \mathcal{S} \rightarrow \mathbb{R}$. An agent acting in an MDP is, at any time step, located at a single state $s \in \mathcal{S}$. The agent chooses an action $a \in \mathcal{A}$ and is relocated to a new state, s', determined by the transition probability distribution $T(s, a, s')$, whereupon it receives reward $R(s')$. In this paper, we are concerned with *goal-seeking* problems in which there is a distinguished "goal state", $g \in \mathcal{S}$, that the agent is trying to reach in the minimum number of steps.

The goal of reinforcement learning in an MDP is to locate a *policy*, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, that specifies an action for the agent in every state. The optimal policy, π^*, is one that maximizes the *value function*, a long-term aggregate measure of received reward. We will use the common *infinite horizon discounted* value function, $V^\pi(s) = \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t)]$, where $0 \leq \gamma < 1$ is the discount factor. In goal-seeking domains, typically a reward of 1 is assigned to g and 0 to all other states and the value function reduces to $V^\pi(s) = \gamma^{\mathrm{E}[\#\text{ steps to reach } g \text{ from } s|\pi]}$.

The MDPs in which we are interested are those derived from a system with an innate topology — specifically, spatial navigational domains. The notion is that, although the agent itself may have stochastic dynamics, there is an underlying topology that governs the deterministic relationships between states. We will assume that there is a distance function on the underlying domain, $d^{topo} : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$. This distance may be interpreted as "the minimum time required to transition between s and s' if all actions were deterministic". We also assume the existence of a coordinate frame in which directions between states are well defined — for example, s' is "northwest" of s — given by a function $\phi^{topo} : \mathcal{S} \times \mathcal{S} \rightarrow \Phi$, where Φ denotes the set of allowable directions such as $\Phi_{\text{Euclidean}} = (-\pi, \pi]$ or $\Phi_{\text{gridworld}} = \mathbb{Z} \times \mathbb{Z}$.[1] Both distance and direction between states are summarized by the *topological relation*, $\mathcal{R}^{topo}_{d,\phi}(s, s')$, between pairs of states. The agent's representation is the set of all possible such equivalence relations, $\mathcal{R} = \{\mathcal{R}^{topo}_{d,\phi}\}$, such that (s_1, s'_1) and $(s_2, s'_2) \in \mathcal{R}^{topo}_{d,\phi} \Leftrightarrow d^{topo}(s_1, s'_1) = d^{topo}(s_2, s'_2)$ and $\phi^{topo}(s_1, s'_1) = \phi^{topo}(s_2, s'_2)$.

This framework allows us to describe useful properties of the MDP in terms of the topology of the underlying domain:

Definition 1 *An MDP is said to be* k-local *iff there is no chance of making a single-step transition between any pair of states further than k units apart with respect to d^{topo}. That is, for every pair of states, s_1, $s_2 \in \mathcal{S}$ such that $d^{topo}(s_1, s_2) > k$, it is the case that $T(s_1, a, s_2) = 0 \; \forall a \in \mathcal{A}$.*

Definition 2 *The* neighborhood *of a state, $\mathcal{N}(s)$, is the set of states reachable from s in a single step with non-zero probability under some action. That is, $\mathcal{N}(s) = \{s' \in \mathcal{S} : \exists a \in \mathcal{A}$ such that $T(s, a, s') > 0\}$.*

In a k-local MDP, all neighborhoods fall within balls of radius k in the underlying topology.

Definition 3 *Two states, s_1 and $s_2 \in \mathcal{S}$ are said to be* isomorphic *iff there is a mapping between their neighborhoods that preserves transition probabilities. That is, s_1 and s_2 are isomorphic iff $\exists f : \mathcal{N}(s_1) \leftrightarrow \mathcal{N}(s_2)$ such that $\forall a \in \mathcal{A}$ and $s'_1 \in \mathcal{N}(s_1)$, $T(s_1, a, s'_1) = T(s_2, a, f(s'_1))$.*

Definition 4 *A subset of the states of an MDP, $S \subseteq \mathcal{S}$, is said to be* homogeneous *iff all states in S are isomorphic according to the above definition.*

Finally, we are specifically interested in systems in which the agent's actions, while noisy, are "largely predictable". That is, there exist actions whose maximum probability outcome is to move the agent in a fixed direction with respect to the underlying topology.

Definition 5 *An action $a \in \mathcal{A}$ is* predictable in S *if, for every state s in a homogeneous subset of the state space, $S \subseteq \mathcal{S}$, a has an outcome state s' with probability $p > 1/2$ having a fixed topological relation to s. That is, $\forall s \in S \; \exists s' \in \mathcal{N}(s)$ such that $T(s, a, s') = p$ and $(s, s') \in \mathcal{R}^{topo}_{d,\phi}$ for some d^{topo} and ϕ^{topo}.*

We call the high-probability outcome of a predictable action the *intended* outcome and any other, lower probability outcomes the *unintended* or *accidental* outcomes. Accidental outcomes may also include any off-policy exploratory actions that the agent takes, so long as the intended outcomes of on-policy actions still occur with probability$> 1/2$.

Relational Policies for Navigation Domains

The traditional definition of policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, is tied to the absolute coordinates of the state space. We would

[1] These are essentially two different agent-centric 2-D polar coordinate systems. Other coordinate systems are also possible, so long as they uniquely describe the relationship between pairs of states.

rather employ policies that are described purely in terms of the topological relationships of the underlying world: $\pi : \mathcal{R}^{topo}_{d,\phi} \rightarrow \mathcal{A}$. Essentially, we are seeking policies of the form "whenever you find yourself X distance southwest of the current goal, act as follows…". Doing so provides two key benefits. First, this representation allows goal-seeking policies to be relocatable, as they now depend only on the relationship between the agent's current location and the goal: when the goal location changes, the agent's policy is automatically defined relative to that new location. This is a form of translation invariance for Markov decision processes. And second, it may allow much faster learning, as relations are entire equivalence classes, so a unit of experience from any element of $\mathcal{R}^{topo}_{d,\phi}$ applies to *all* elements of that class. This is, however, profitable only when there may be multiple goal states whose absolute locations vary over time or when it is useful to formulate an overall policy in terms of a number of sub-goals. This is a plausible condition; there are numerous real-world tasks that require seeking a number of different locations within the same environment, and many previous authors have demonstrated that sub-goals are useful for planning and learning in MDPs (Parr 1998b; Dietterich 2000; Precup 2000; Lane & Kaelbling 2002).

The utility of reasoning with such relationships has been well established for deterministic domains. Unfortunately, things are more complicated in stochastic domains. In a deterministic environment, we can assume that any desired trajectory can be followed exactly and that no off-trajectory obstacles can interfere. In a stochastic MDP, however, exact analysis of the transition between any pair of states involves consideration of *all* trajectories between them, and even obstacles that are not on the "optimal" trajectory can still influence the expected transition time. Such obstacles can even produce asymmetries in expected transition times, which prevents us from using the transition time directly as a metric.

In this section, we demonstrate the utility of reinforcement learning with relational policies in a trivial environment in which exact relocatability is achievable. We then consider environments containing obstacles and demonstrate that, so long as the obstacles are "far" from an optimal trajectory, that their influence can be safely neglected, allowing relocation of policies in less trivial environments.

The Exact Case

We begin with an extremely simple domain topology: a perfectly homogeneous toroidal gridworld. In this domain, every state s has exactly four neighbors, denoted NORTH(s), SOUTH(s), EAST(s), and WEST(s). The topology has the intuitive interpretation so that, e.g., NORTH(EAST(s))=EAST(NORTH(s)), etc. The agent has four predictable actions available to it, corresponding to the four neighbor states. Each action has its intended outcome with probability $p = 0.9$ while its unintended outcomes place the agent at each of the other three neighbor states with probability $(1-p)/3 \approx 0.03$. The torus has a circumference of 50 states in the east/west direction and 20 states in the north/south direction. All states are isomorphic according to Definition 3, implying that there are no walls, cliffs, absorbing states, or other obstacles in the environment. We define $d^{topo}(s, s')$ to be the minimum Manhattan distance between s and s' and $\phi^{topo}(s, s')$ to be the ordered pair (dx, dy) representing the ratio of horizontal to vertical change between s and s', reduced to lowest terms. The agent learns a Q function expressed in terms of this relational representation, $Q : \mathcal{R} \times \mathcal{A} \rightarrow \mathbb{R}$, allowing relocation of policies as goal locations change.

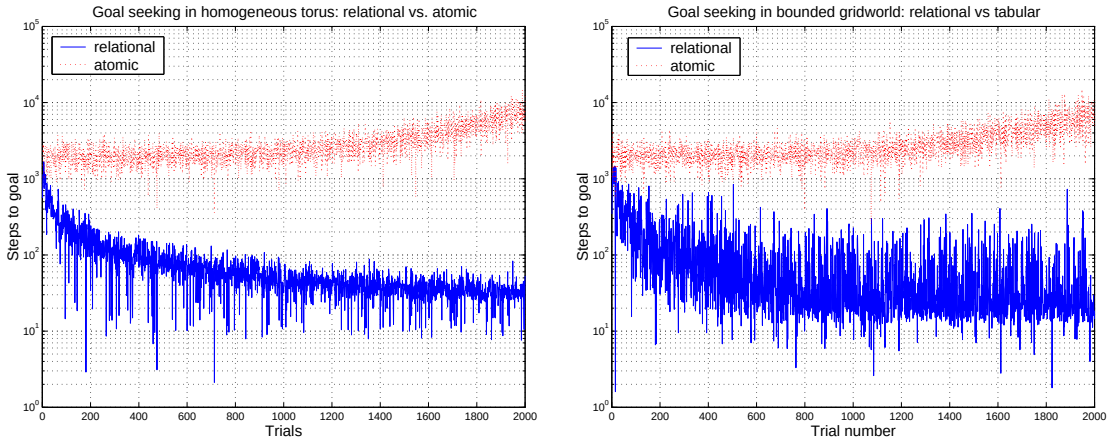
Ravindran and Barto (Ravindran & Barto 2002; 2003; Ravindran 2004), extending model minimization work by Dean and Givan (Dean & Givan 1997), have studied such policy relocations. Ravindran and Barto have developed a very general algebraic description of policy relocations in terms of homomorphisms between MDPs. In the case of the homogeneous torus, all spatial translations of the coordinate system are isomorphic (a special case of their more general homomorphisms) and their results imply that policies are exactly relocatable through translations. In our case, this means that reinforcement learning in terms of our relational Q functions/policies should be exact and (assuming a convergent learning method) should converge to an optimal policy.

We examined the performance of two learning agents, one employing an atomic state representation, the other a relational representation, in this environment. Both agents employed Q-learning with ϵ-greedy exploration policy (Sutton & Barto 1998) with a discount factor of $\gamma = 0.99$, a learning rate of $\alpha = 0.7$, and an exploration factor of $\epsilon = 0.01$. We ran each agent for 2000 trials in this world, where each trial consisted of selecting a random start state and a random goal location and allowing the agent to run until it encountered the goal. We recorded the number of steps that the agent took to locate the goal on each trial. We repeated the entire operation 10 times and averaged the results over all 10 runs. While this environment is stationary for the relational representation, it is non-stationary for the atomic agent and we do not expect the atomic agent to converge to a stable policy. We also attempted an agent that used an atomic representation including a variable for the current location of the goal in addition to the variable giving the agent's location (total of $O(N^2)$ states). This representation renders the environment stationary, but it was so large (sim30 Mb) that the agent was unable to make any measurable progress in the amount of time that we could dedicate to testing it.

Figure 1 (**a**) shows the learning performance of the relational and atomic agents. As expected, the atomic agent is not able to learn a stable policy in this world (in fact, its performance diverges). The relational agent, however, does converge to a stable, high-quality solution fairly quickly. By convergence, the relational agent had located a policy that achieved a mean time-to-goal of 25.1 steps — close to the expected transition time for a randomly chosen pair of states of about 22 steps.

Limited Non-Homogeneity

Homogeneous tori are not terribly useful, however — interesting navigational environments include non-homogeneous features such as walls, cliffs, or obstacles. These features



(**a**) (**b**)

Figure 1: (**a**) Performance of atomic (upper trace) and relational (lower trace) Q-learning agents in the homogeneous torus gridworld. Both curves are averages over ten different sequences of (start,goal) pairs in the torus. (**b**) The same experiment repeated for non-toroidal gridworld.

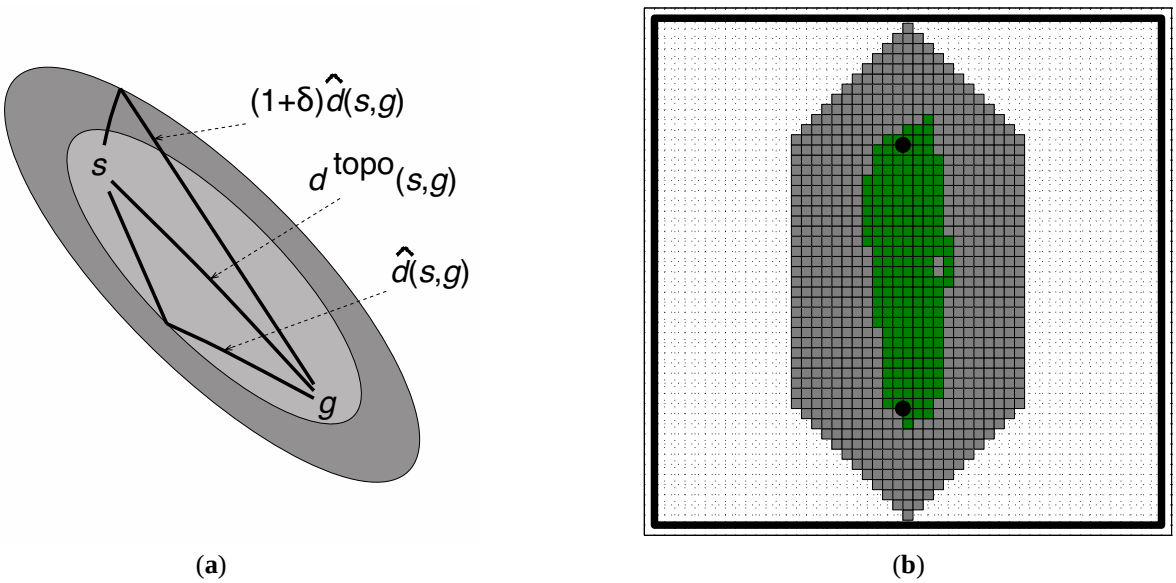
render policy relocation problematic because they change the set of trajectories (and the probability of each) that an agent could experience in attempting to transition from s to g. Even if a wall, say, does not obstruct the shortest path between s and g, it may block a trajectory that an agent would otherwise experience.

In general this problem is difficult to overcome. In ongoing work, we are examining the possibility of adding additional relational terms to describe the relative position of walls/obstacles with respect to the agent and goal. In this paper, however, we demonstrate a weaker, but still useful, approach. Specifically, if an agent is able to find a reasonably good policy for navigating to a goal in one part of the space, it can relocate that policy to a different part of the space with small expected change in value, so long as any such obstacles are "far" from the ideal path to the goal. We give an envelope that contains at least $1-\epsilon$ of the probability mass of all trajectories that the agent might experience in transitioning from s to g. Assuming that the MDP has rewards bounded by $R_{\max} < \infty$, the presence of walls or obstacles outside the envelope can change the agent's expected value by at most $\epsilon R_{\max}$.

We obtain a bound on the envelope of probability $(1-\epsilon)$ as follows. Assume that the agent is executing in a k-local MDP that is "largely" homogeneous — that is, the majority of the state space is homogeneous, with the exception of the existence of some walls or other obstacles. Consider an agent starting at state s attempting to reach state g according to a fixed policy π. Let the actual *expected* transition time between s and g be $\hat{d}(s,g)$, which is a function of s, g, and π. π need not be an optimal policy to reach g, so long as $\hat{d}(s,g)$ is within a constant factor of $d^{topo}(s,g)$. Note that the set of all states reachable by a trajectory of length $\hat{d}(s,g)$ that starts at s and ends at g forms an ellipse with respect to d^{topo}: for any s' along the trajectory, it must be the case that $d^{topo}(s,s') + d^{topo}(s',g) \leq k\hat{d}(s,g)$ by virtue of k-locality and the metric of the underlying space. The major axis of this ellipse lies along the shortest path between s and g according to d^{topo}.

Without loss of generality, assume that an intentional outcome reduces $d^{topo}(agent,g)$ by at least 1 unit at every state. Some lower bound is guaranteed by the assumption of transition time and we can rescale d^{topo} as necessary. By k-locality, we know that any accidental outcome can increase $d^{topo}(agent,g)$ by at most k units. Thus, in expectation, the agent moves $p-k(1-p) = (k+1)p-k$ units toward the goal every time step. When this quantity is positive, the agent will reach the goal in $\hat{d}(s,g) = d^{topo}(s,g)/((k+1)p-k)$ steps.

Homogeneity allows us to model the sequence of intentional/unintentional outcomes as a series of Bernoulli trials. The number of actions necessary for the agent to reach the goal is given by the negative binomial distribution: the number of accidental outcomes between any pair of intentional outcomes is geometrically distributed and the total transition time is given by a sum of geometrically distributed variables. A Chernoff bound assures us that the probability that this sum deviates far from its mean is exponentially small: $\Pr[|\text{trajectory}| > (1+\delta)\hat{d}(s,g)] < \exp(-\hat{d}(s,g)\delta^2/4)$. That is, the chance that the agent will take significantly longer than $\hat{d}(s,g)$ to reach the goal falls off exponentially with δ. As we have argued above, because of the strong constraints of the underlying topology, any trajectory of length $(1+\delta)\hat{d}(s,g)$ must fall within an ellipse surrounding the optimal path from s to g. To ensure that this elliptical envelope contains at least $(1-\epsilon)$ of the probability mass, we take $\delta > \sqrt{\frac{-4\ln\epsilon}{\hat{d}(s,g)}}$. Figure 2 gives an intuitive view of such an envelope. This is actually a fairly loose bound, as it assumes that every accidental move is maximally detrimental to the agent. In practice, many agents make symmetric errors, so accidents can be somewhat self-compensating. The elliptical form of the envelope also assumes that all accidental outcomes could occur in a sequence, carrying the agent as far as possible from g — also a low probability circumstance. We are currently exploring how to tighten this envelope by



(**a**) (**b**)

Figure 2: (**a**) Schematic diagram of envelope bound structure. The inner, light gray ellipse is the set of trajectories of length $\leq \hat{d}(s,g)$; the outer, dark gray ellipse is the envelope of probability $(1-\epsilon)$. With high probability, no trajectory between s and g under a reasonably good policy will leave the outer ellipse. (**b**) Empirical examination of the envelope in a homogeneous gridworld. The start and goal states are indicated by dark circles; the dark, inner filled squares are states encountered by the agent in 10,000 steps of experience (319 trajectories). The outer, paler filled squares (hexagon) are the ellipse of the $(1-0.1)$ envelope under the Manhattan metric.

exploiting such properties.

The implication of this result is that all homogeneous, elliptical regions of an MDP obeying the above constraint are isomorphic and that goal-seeking policies executed within those regions are value equivalent up to a factor of $\epsilon R_{\max}$. So, for example, if an agent has acquired a reasonably effective policy for transitioning between states s and g, having relationship $\mathcal{R}^{topo}_{d,\phi}(s,g)$, then that policy can be relocated to any other pair of states $(s',g') \in \mathcal{R}^{topo}_{d,\phi}(s,g)$ so long as s' and g' belong to a homogeneous elliptical envelope.

While this result doesn't apply to the entire training lifetime of an RL agent, it does apply once the agent has located an initial, reasonably good policy. Thereafter, these envelopes are tolerant of exploratory actions and policy improvements will only tighten the envelope.

We repeated the "Exact Case" experiments for a similar but non-toroidal gridworld. This world is also largely homogeneous, except for the presence of outer walls that prevent moves from wrapping around. In this environment, most (*start,goal*) pairs do obey the ellipse constraints specified above, except for those very near the walls. We constructed a relational agent using pure Manhattan distance for d^{topo} (as opposed to the toroidal Manhattan distance that we used previously) and tested it against an atomic agent in this environment. The results, given in Figure 1 (**b**), display the same learning patterns as those in (**a**). The small amount of inhomogeneity introduced by adding walls does not significantly degrade the performance of the relational learning agent, though it does increase the variance. The agent *does* however, require built-in knowledge of d^{topo} — when we trained a relational agent using the toroidal Manhattan distance on the non-toroidal gridworld, its policy quickly diverged and it didn't even complete the 2000 trials in the allotted 10^7 steps.

Conclusion

We have argued that stochastic navigational domains possess important topological structure that we can exploit to build efficient relational reinforcement learning agents. We described a relational policy representation that exploits knowledge of the underlying topology and showed that it can be used to substantial advantage in simple, open space navigation domains. We identified important characteristics of an MDP — locality, homogeneity, and action predictability — and described them in terms that relate the MDP to the underlying topology. We used these three properties to derive a closed-form bound for an envelope, given a reasonable goal-seeking policy.

This work is clearly only a first step toward a widely applicable theory of topologically constrained navigational reinforcement learning. In this work, we have only shown the utility of relational RL in largely "open space" environments. Our envelope bound allows nonhomogeneities *outside* the bounded region, but still requires the bounded region to be perfectly homogeneous and unobstructed – conditions of admittedly limited practical significance. Nonetheless, our observations represent a key first step as they are, to our knowledge, the first use of metric and topology to constrain the general RL case. Not only do these properties allow us to *a priori* discard impossible conditions (e.g., tele-

porting across the world), but they also allow us to quickly ascertain and neglect *highly improbable* conditions (such as violating the elliptical envelope bound). We can thus realize two learning speedups: we can employ simple models of transition functions that can be learned more efficiently than can general multinomial models, and we can generalize experience across isomorphic envelopes. Generalization of experience across different MDPs, or across different regions of a single MDP, is one of the holy grails of reinforcement learning, and we believe that exploiting properties such as metric or topology gives us a principled framework in which to do so for specific classes of MDPs.

In current work we are extending the insights of this paper to more general navigational problems. We employ a more general relational description language that allows descriptions of the form "`Here(`s_0`)` AND `Wall(`s_1`)` AND `NorthEast`$_{3,5}$`(`s_1`,`s_0`)` AND `TerrainMud(`s_0`)` AND ...". This language encodes both metric, local topology, and other features of the world that are relevant to motion dynamics. Using bounds on biased random walks, we are extending the envelope results of this paper to the more general relational framework, allowing policy generalization across a much richer class of navigational environments. Interestingly, the use of strong envelope bounds can be viewed as a certain sort of partially-observable (POMDP) planning – only states within the envelope are relevant to the agent, so unobservability of states outside the envelope will not change the agent's actions. Finally, we are working to couple the current approach to our previous results on mixed (MDP and deterministic graph) planning in "nearly deterministic" environments (Lane & Kaelbling 2002). Our ultimate vision is a multi-level learning and planning system, in which RL is responsible for learning local structure, transition dynamics, noise models, and control, while high-level task planning and goal prioritization can be left to a deterministic planner.

References

Anderson, C. R.; Domingos, P.; and Weld, D. S. 2002. Relational Markov models and their application to adaptive web navigation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*. Edmonton, Alberta, Canada: ACM SIGKDD.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 2000. Stochastic dynamic programming with factored representations. *Artificial Intelligence* 121(1–2):49–107.

Dean, T., and Givan, R. 1997. Model minimization in Markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, 106–111. Providence, RI: AAAI Press/MIT Press.

Dean, T.; Kaelbling, L. P.; Kirman, J.; and Nicholson, A. 1995. Planning under time constraints in stochastic domains. *Artificial Intelligence* 76.

Dietterich, T. G. 2000. An overview of MAXQ hierarchical reinforcement learning. In Choueiry, B. Y., and Walsh, T., eds., *Proceedings of the Symposium on Abstraction, Reformulation and Approximation (SARA 2000)*, Lecture Notes in Artificial Intelligence. New York: Springer Verlag.

Finney, S.; Gardiol, N. H.; Kaelbling, L. P.; and Oates, T. 2002. The thing that we tried didn't work very well: Deictic representation in reinforcement learning. In *Proceedings of the Eighteenth International Conference on Uncertainty in Artificial Intelligence (UAI-2002)*.

Getoor, L.; Friedman, N.; Koller, D.; and Pfeffer, A. 2001. Learning Probabilistic Relational Models. In Dzeroski, S. and Lavrac, N., eds., *Relational Data Mining*. Springer-Verlag.

Hauskrecht, M.; Meuleau, N.; Boutilier, C.; Kaelbling, L. P.; and Dean, T. 1998. Hierarchical solution of Markov decision processes using macro-actions. In Cooper, G. F., and Moral, S., eds., *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*. Morgan Kaufmann.

Lane, T., and Kaelbling, L. P. 2002. Nearly deterministic abstractions of Markov decision processes. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, 260–266. Edmonton, Canada: AAAI Press.

Parr, R. 1998a. Flexible decomposition algorithms for weakly coupled Markov decision problems. In Cooper, G. F., and Moral, S., eds., *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*. Morgan Kaufmann.

Parr, R. E. 1998b. *Hierarchical Control and Learning for Markov Decision Processes*. Ph.D. Dissertation, University of California at Berkeley.

Precup, D. 2000. *Temporal Abstraction in Reinforcement Learning*. Ph.D. Dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons.

Ravindran, B., and Barto, A. G. 2002. Model minimization in hierarchical reinforcement learning. In Holte, R., ed., *Proceedings of the 2002 Symposium on Abstraction, Reformulation, and Approximation (SARA-2002)*.

Ravindran, B., and Barto, A. G. 2003. Relativized options: Choosing the right transformation. In Fawcett, T., and Mishra, N., eds., *Proceedings of the Twentieth International Conference on Machine Learning*, 608–615. Washington, DC: AAAI Press.

Ravindran, B. 2004. *An Algebraic Approach to Abstraction in Reinforcement Learning*. Ph.D. Dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.