In *The Fourth International Joint Conference on Autonomous Agents and Multiagent Systems* (AAMAS-05), pp. 53-59, Utrecht, The Netherlands, July 2005.

Behavior Transfer for Value-Function-Based Reinforcement Learning

Matthew E. Taylor and Peter Stone
Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712-1188
{mtaylor, pstone}@cs.utexas.edu

ABSTRACT

Temporal difference (TD) learning methods [22] have become popular reinforcement learning techniques in recent years. TD methods have had some experimental successes and have been shown to exhibit some desirable properties in theory, but have often been found very slow in practice. A key feature of TD methods is that they represent policies in terms of value functions. In this paper we introduce *behavior transfer*, a novel approach to speeding up TD learning by transferring the learned value function from one task to a second related task. We present experimental results showing that autonomous learners are able to learn one multiagent task and then use behavior transfer to markedly reduce the total training time for a more complex task.

1. INTRODUCTION

Temporal difference learning methods [22] are a type of reinforcement learning that has shown some success in different machine learning tasks because of its ability to learn where there is limited prior knowledge and minimal environmental feedback. However, TD often is very slow in practice to produce near-optimal behaviors. Many techniques exist which attempt, with more or less success, to speed up the learning process.

Two generally accepted techniques for solving reinforcement learning problems are TD methods and policy search methods [23]. Direct policy search methods modify parameters of a policy, a way of acting in a particular task, so that over time the resulting policy will improve. Agents using policy search methods explore policy space by adjusting parameters of the policy and then observing the resulting performance. In contrast, value-based methods learn to estimate a *value function* for each situation that the learner could find itself in. The learner is then able to take the action which it believes will give it the most value in the long run. Over time the learned value function approaches the true value of each state by comparing the expected value of a state with the actual value received in that state. Over time the agent will build an accurate measure of how valuable each available action is. Both policy search and value-based methods are able to determine the optimal policy but only value-based methods are inherently able to provide an estimate of the "goodness" of taking a given action from a given state.

Past research [17] has shown that a learner can train faster on a task if it has first learned on a simpler variation of the task, referred to as *directed training*. In this paradigm the state transition function, which is part of the environment, can change between tasks. *Learning from easy missions* [4] is a technique that relies on human input to modify the starting state of the learner over time, making it incrementally more difficult for the learner. Both of these methods reduce the total training time required to successfully learn the final task. However, neither allow for changes to the state or action spaces between the tasks, limiting their applicability. *Reward shaping* [6, 11] allows one to bias a learner's progress through the state space by adding in artificial rewards to the environmental rewards. Doing so requires sufficient knowledge about the environment a priori to guide the learner and must be done carefully to ensure that unintended behaviors are not introduced. While it is well understood how to add this type of guidance to a learner [13], we would prefer to allow the agent to learn faster by training on different (perhaps pre-existing) tasks rather than creating artificial tasks that are easier to learn.

In this paper we introduce *behavior transfer*, whereby a TD learner trained on one task can learn faster when training on another task with related, but different, state and action spaces. Behavior transfer is more general than the previously referenced methods because it does not preclude the modification of the transition function, start state, or reward function. The key technical challenge is mapping a value function in one representation to a meaningful value function in another, typically larger, representation. The primary contribution of this paper is to establish an existence proof that there are domains in which it is possible to construct such a mapping and thereby speed up learning via behavior transfer. While we expect that behavior transfer is possible for both TD and policy search learners, the two types of learners will likely require different techniques; in this paper we address only behavior transfer in TD learners.

The remainder of this paper is organized as follows. Section 2 formally defines the behavior transfer methodology and its benefits. Section 3 gives an overview of the domain in which we quantitatively test behavior transfer. Section 4 gives details of learning in this domain. Section 5 describes how we perform behavior transfer in the test domain. Section 6 presents the results of our experiments and discusses some of their implications. Section 7 details other related work while contrasting our methods and Section 8 concludes.

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

2. BEHAVIOR TRANSFER METHODOLOGY

To formally define behavior transfer we first briefly review the general reinforcement learning framework that conforms to the generally accepted notation for Markov decision processes [15]. There is some set of possible perceptions of the current state of the world, S, and a learner has some initial starting state, $s_{initial}$. When in a particular state s there is a set of actions A which can be taken. The reward function R maps each perceived state of the environment to a single number which is the value, or instantaneous reward, of the state. The transition function, T, takes a state and an action and returns the state of the environment after the action is performed. If transitions are non-deterministic the transition function is a probability distribution function. A learner is able to sense s, and typically knows A, but may or may not initially know S, R, or T.

A policy $\pi : S \mapsto A$ defines how a learner interacts with the environment by mapping perceived environmental states to actions. π is modified by the learner over time to improve performance, i.e. the expected total reward accumulated, and it completely defines the behavior of the learner in an environment. In the general case the policy can be stochastic. The success of an agent is determined by how well it maximizes the total reward it receives in the long run while acting under some policy π. An *optimal policy*, π^*, is a policy which does maximize this value (in expectation). Any reasonable learning algorithm attempts to modify π over time so that it reaches π^* in the limit.

Past research confirms that if two tasks are closely related the learned policy from one task can be used to provide a good initial policy for the second task. For example, Selfridge (1985) showed that the 1-D pole balancing task could be made harder over time by shortening the length of the pole and decreasing its mass; when the learner was first trained on a longer and lighter pole it could more quickly learn to succeed in the more difficult task with the modified transition function. In this way, the learner is able to refine an initial policy for a given task: $(S_1, s_{(1,initial)}, A_1, T_1, R_1, \pi_0) \Rightarrow \pi_{(1,final)}$ where task 1 starts from no initial policy as indicated by the π_0 in the last value of the tuple. Task 2 can then be defined as $(S_2, s_{(2,initial)}, A_2, T_2, R_2, \pi_{(2,initial)}) \Rightarrow \pi_{(2,final)}$. The time it takes to learn $\pi_{(2,final)} = \pi_2^*$ may be less for $(S_2, s_{(2,initial)}, A_2, T_2, R_2, \pi_{(1,final)})$ than $(S_2, s_{(2,initial)}, A_2, T_2, R_2, \pi_0)$. Note that since $S_1 = S_2$ and $A_1 = A_2$, $\pi_{(1,final)}$ is a legitimate policy for task 2.

In this paper we consider the more general case where $S_1 \neq S_2$, and/or $A_1 \neq A_2$. To use the policy $\pi_{(1,final)}$ as the initial policy for a TD learner in second task, we must transform the value function so that it can be directly applied to the new state and action space. A behavior transfer functional $\rho(\pi)$ will allow us to apply a policy in a new task $(S_2, s_{(2,initial)}, A_2, T_2, R_2, \rho(\pi_{(1,final)})) \Rightarrow \pi_{(2,final)}$. The policy transform functional, ρ, needs to modify the policy and its associated value function so that it accepts the states in the new task as inputs and allows for the actions in the new task to be outputs, as depicted in Figure 1. A policy generally selects the action which is expected to accumulate the largest expected total reward and thus the problem of transforming a policy between two tasks reduces to transforming the value function. In this paper we will therefore concentrate on transferring the state action values, Q, from one learner to another. Defining ρ to do this correctly is the key technical challenge to enable general behavior transfer.

One measure of success in speeding up learning using this method is that given a policy π_1, the training time for $\pi_{(2,final)}$ to reach some performance threshold decreases when replacing the initial policy π_0 with $\rho(\pi_1)$. This criterion is relevant when task 1 is given and is of interest in its own right. A stronger measure of success is that the training time for both tasks using behavior transfer is shorter than the training time to learn the second task from scratch. This criterion is relevant when task 1 is created for the sole purpose of speeding up learning via behavior transfer.

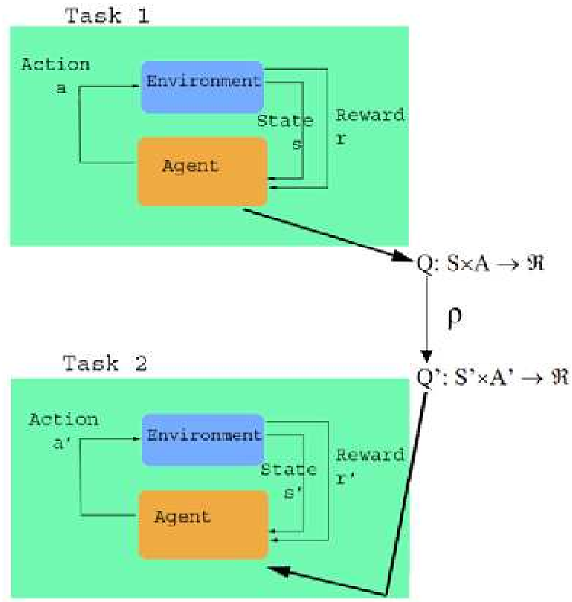

Figure 1: ρ is a functional which transforms a value function Q from one task so that it is applicable in a second task that has different state and action spaces.

3. TESTBED DOMAIN

To demonstrate the effectiveness and applicability of the behavior transfer method (detailed in section 5) we empirically test it in the RoboCup simulated soccer keepaway domain using a setup similar to past research [21]. RoboCup simulated soccer is well understood as it has been the basis of multiple international competitions and research challenges. The multiagent domain incorporates noisy sensors and actuators, as well as enforcing a hidden state so that agents can only have a partial world view at any given time. While there has been previous work which attempted to use machine learning to learn the full simulated soccer problem [3, 16], the complexity and size of the problem have so far proven prohibitive. However, many of the RoboCup subproblems have been isolated and solved using machine learning techniques, including the task of playing keepaway.

Keepaway, a subproblem of RoboCup soccer, is the challenge where one team, the *keepers*, attempts to maintain possession of the ball on a field while another team, the *takers*, attempts to gain possession of the ball or force the ball out of bounds, ending an *episode*. Keepers that are able to make better decisions about their actions are able to maintain possession of the ball longer and thus have a longer average episode length. Figure 2 depicts three keepers playing against two takers. [1]

As more players are added to the task, keepaway becomes harder for the keepers because the field becomes more crowded. As more takers are added there are more players to block passing lanes and chase down any errant passes. As more keepers are added, the

[1]Flash-file demonstrations of the task can be found at *http://www.cs.utexas.edu/users/AustinVilla/sim/keepaway/*.

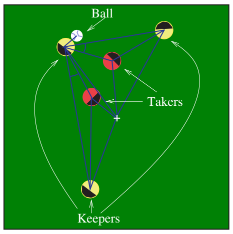

Figure 2: This diagram depicts the 13 state variables used for learning with 3 keepers and 2 takers. There are 11 distances to players, the center of the field, and the ball, as well as 2 angles along passing lanes.

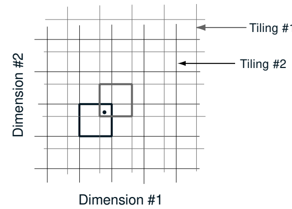

Figure 3: The tile-coding feature sets are formed from multiple overlapping tilings. The state variables are used to determine the activated tile in each of the different tilings. Every activated tile then contributes a weighted value to the total output of the CMAC for the given state. Increasing the number of tilings allows the tile-coding to generalize better while decreasing the tile size allows more accurate representations of smaller details. Note that we primarily use one-dimensional tilings but that the principles apply in the n-dimensional case.

keeper with the ball has more passing options but the average pass distance is shorter. This forces more passes and will lead to more errors because of the noisy actuators and imperfect perception. For this reason keepers in 4 vs. 3 keepaway (meaning 4 keepers and 3 takers) take longer to learn an optimal control policy than in 3 vs. 2. The hold time of the best policy for a constant field size will also decrease when moving from 3 vs. 2 to 4 vs. 3 due to the added difficulty. This has been discussed in previous research [21]. The time it takes to reach a policy which is near the handcoded solution roughly doubles as each additional keeper and taker is added.

The different keepaway tasks are all situations which may occur during a real game. Learning on one task and transferring the behavior to a separate useful task can reduce the training time. In the keepaway domain, A and S are determined by the current keepaway task and thus differ from instance to instance. However, $s_{initial}$, R, and T, though formally different, are effectively constant across tasks. When S and A change, $s_{initial}$, R, and T change by definition. But in practice, R is always defined as 1 for every time step that the keepers maintain possession, and $s_{initial}$ and T are always defined by the RoboCup soccer simulation.

4. LEARNING KEEPAWAY

The keepers use episodic SMDP Sarsa(λ) [22], a well understood temporal difference algorithm, to learn their task. We use linear tile-coding function approximation, also known as CMACs, which has been successfully used in many reinforcement learning systems [1]. The keepers choose not from primitive actions (turn, dash, or kick) but higher-level actions first implemented by the CMUnited-99 team [20]. A keeper without the ball automatically attempts to move to an open area (the receive action). A keeper in possession of the ball has the freedom to decide whether to hold the ball or to pass to a teammate.

Our agents are based on the keepaway benchmark players distributed by UT-Austin[2] which are described in [19]. These benchmark players are built on the UvA Trilearn team [5] and the CMUnited-99 team [20], whereas previous publications [21] use players built on the CMUnited-99 players. The newer benchmark players have better low-level functionality and are thus able to hold the ball for longer than the older CMUnited-99 players, both before and after learning.

Function approximation is often needed in reinforcement learning so that the learner is capable of generalizing the policy to perform well on unvisited states. CMACs allow us to take arbitrary groups of continuous state variables and lay infinite, axis-parallel tilings over them (see Figure 3). Using this method we are able to discretize the continuous state space by using tilings while maintaining the capability to generalize via multiple overlapping tilings. The number of tiles and width of the tilings are hardcoded and this dictates which state values will activate which tiles. The function approximation is learned by changing how much each tile contributes to the output of the function approximator. By default, all the CMAC's weights are initialized to zero. This approach to function approximation in the RoboCup soccer domain is detailed by Stone and Sutton (2002).

For the purposes of this paper, it is particularly important to note the state variables and action possibilities used by the learners. The keepers' states comprise distances and angles of the keepers $K_1 - K_n$, the takers $T_1 - T_m$, and the center of the playing region C (see Figure 2). Keepers and takers are ordered by increasing distance from the ball. Note that as the number of keepers *n* and the number of takers *m* increase, the number of state variables also increase so that the more complex state can be fully described. S must change (e.g. there are more distances to players to account for) and $|A|$ increases as there are more teammates for the keeper with possession of the ball to pass to. Full details of the keepaway domain and player implementation are documented elsewhere [21].

4.1 Learning 3 vs. 2

On a 25m x 25m field, three keepers are initially placed in the three corners of the field and a ball is placed near one of the keepers. The two takers are placed in the fourth corner. When the episode starts, the three keepers attempt to keep control of the ball by passing amongst themselves and moving to open positions. The keeper with the ball has the option to either pass the ball to one of its two teammates or to hold the ball. We allow the keepers to learn to choose between these three choices when in control of the ball. In this task A = {hold, passToTeammate1, passToTeammate2}. S is defined by 13 state variables, as shown in Figure 2. When a taker gains control of the ball or the ball is kicked out of the field's bounds the episode is finished. The reward to the Sarsa(λ) algorithm for the keeper is the number of time steps the ball remains in play after an action is taken. The episode is then reset with a random keeper placed near the ball.

All weights in the CMAC function approximator are initially set to zero and therefore $\pi_{(3vs2,initial)} = \pi_0$. As training progresses, the weight values are changed by Sarsa(λ) so that the average hold time of the keepers increases. Throughout this process, the takers use a static hand-coded policy to attempt to capture the ball as quickly as possible. Due to the large amounts of randomness in the

[2]Source code, doumentation, and mailing list can be found at http://www.cs.utexas.edu/users/AustinVilla/sim/keepaway/.

environment, the evaluation of a policy is very noisy.

4.2 Learning 4 vs. 3

Holding the field size constant we now add an additional keeper and an additional taker. R and T are essentially unchanged from 3 vs. 2 keepaway, but now A = {hold, passToTeammate1, passToTeammate2, passToTeammate3} and S is made up of 19 state variables due to the added players. The 4 vs. 3 task is harder than the 3 vs. 2 task and the learned average hold times after 20 hours of training from $\pi_{initial} = \pi_0$ decrease from roughly 13.6 seconds for 3 vs. 2 to 9.3 seconds for 4 vs. 3.

In order to quantify how fast an agent in 4 vs. 3 learns, we set a threshold of 9.0 seconds. When a group of four keepers has learned to hold the ball from the three takers for an average of 9.0 seconds over 1,000 episodes we say that the keepers have sufficiently learned the 4 vs. 3 task. By recording this time over many trials we can measure the effectiveness of the Sarsa(λ) algorithm in different situations.

5. BEHAVIOR TRANSFER IN KEEPAWAY

To define a ρ which will correctly transfer behavior from $\pi_{(3vs2,final)}$ into $\pi_{(4vs3,initial)}$ for a TD learner, the value function utilized by π needs to handle the new state and action spaces reasonably. In the keepaway domain we are able to intuit the mappings between actions in the two tasks and states in the two tasks based on our knowledge of the domain. Our choice for the mappings is supported by empirical evidence in Section 6 showing that using this mapping in behavior transfer does decrease training time. Other domains will not necessarily have such straightforward transforms between tasks of different complexity. Finding a general method to specify ρ is outside the scope of this paper and will be formulated in future work. One of the main challenges will be identifying general heuristics for mapping existing states and actions in the first task to new states and actions in a second task. Creating a general metric for similarity between state variables and actions in two tasks would allow us to identify a promising mapping for ρ and give an a priori indication of whether behavior transfer will work in a particular domain. Our primary contribution in this paper is demonstrating that there exist domains in which ρ can be constructed and then used to successfully increase the learning rate.

The naive approach of directly copying the CMAC's array of weights to duplicate the value function from $\pi_{(3vs2,final)}$ into $\pi_{(4vs3,initial)}$ without any adjustment fails because both S and A have changed. Keeping in mind that $\pi : S \mapsto A$, we can see that the new state vectors which describe the learner's environment would not necessarily be correctly used, nor would the new actions be correctly evaluated by $\pi_{(3vs2,final)}$. In order to use the learned policy we modify it to handle the new actions and new state values in the second task so that the CMAC can reasonably evaluate them.

The CMAC function approximator takes a state vector and an action and returns the expected total reward. The learner can evaluate each potential action for the current S and then use π to choose one. We modify the weights in the tile coding so that when we input a 4 vs. 3 action the weights for the activated tiles are not zero but instead are initialized by $\pi_{3vs2,final}$. To accomplish this, we copy weights from the tiles which would be activated for a similar action in 3 vs. 2 into the tiles activated for every new action. The weights corresponding to the tiles that are activated for the "pass to teammate 2" action are copied into the weights for the tiles that are activated to evaluate the "pass to teammate 3" action. The modified CMAC will initially be unable to distinguish between these two actions.

To handle new state variables we follow a similar strategy. The 13 state variables which are present in 3 vs. 2 are already handled by the CMAC's weights. The weights for tiles activated by the six new 4 vs. 3 state variables are initialized to values of weights activated by similar 3 vs. 2 state variables. For instance, weights which correspond to "distance to teammate 2" values in the state representation are copied into the weights for tiles that are used to evaluate "distance to teammate 3" state values. This is done for all six new state variables. As a final step, any weights which have not been initialized are set to the average value of all initialized weights. The 3 vs. 2 training was not exhaustive and therefore some weights which may be utilized in 4 vs. 3 would otherwise remain uninitialized. In this way, the tiles which correspond to every value in the new 4 vs. 3 state vector have been initialized to values determined via training in 3 vs 2 and can therefore be considered in the computation. See Table 1 for examples of mappings used. Identifying similar actions and states between two tasks is essential for constructing ρ and may prove to be the main limitation when attempting to apply behavior transfer to different domains.

Having constructed a ρ which handles the new states and actions, we can now set $\rho(\pi_{(3vs2,final)}) = \pi_{(4vs3,initial)}$. We do not claim that these initial CMAC weights are correct (and empirically they are not), but instead that the constructed CMAC allows the learner to more quickly discover a near-optimal policy.

6. RESULTS AND DISCUSSION

To test the effect of loading the 3 vs. 2 CMAC weights into 4 vs. 3 keepers, we run a number of 3 vs. 2 episodes, save the CMAC weights ($\pi_{(3vs2,final)}$) from a random 3 vs. 2 keeper, and load the CMAC weights into all four keepers[3] in 4 vs. 3 so that $\rho(\pi_{(3vs2,final)}) = \pi_{(4vs3,initial)}$. Then we train on the 4 vs. 3 keepaway task until the average hold time for 1,000 episodes is greater than 9.0 seconds. To overcome the high variance inherent in the environment and therefore the noise in our evaluation, we run at least 100 independent trials for each number of 3 vs. 2 training episodes.

Table 2 reports the average time spent training 4 vs. 3 to achieve a 9.0 second average hold time for different amounts of 3 vs. 2 training. The middle column reports the time spent training on the 4 vs. 3 task while the third column shows the total time taken to train 3 vs. 2 and 4 vs. 3. As can be seen from the table, spending time training in the simpler 3 vs. 2 domain can cause the time spent in 4 vs. 3 to decrease.

Table 2 shows the potential of behavior transfer. We use a t-test to determine that the differences in the distributions of 4 vs. 3 training times when using behavior transfer are statistically significant ($p < 5.7 * 10^{-11}$) when compared to training 4 vs. 3 from scratch. Not only is the time to train the 4 vs. 3 task decreased when we first train on 3 vs. 2, but the total training time is less than the time to train 4 vs. 3 from scratch. We can therefore conclude that in the keepaway domain training first on a simpler task can increase the rate of learning enough that the total training time is decreased. To verify that the 4 vs. 3 players were benefiting from behavior transfer and not from simply having non-zero initial weights, we initialized the 4 vs. 3 CMAC weights uniformly to 1.0 in one set of experiments and then to random numbers from 0.0-0.5 in a second

[3]We do so under the hypothesis that the policy of a single keeper represents all of the keepers' learned knowledge. Though in theory the keepers could be learning different policies that interact well with one another, so far there is no evidence that they do. One pressure against such specialization is that the keepers' start positions are randomized. In earlier informal experiments, there appeared to be some specialization when each keeper started in the same location every episode.

4 vs. 3 state variable	related 3 vs. 2 state variable
$dist(K_1, C)$	$dist(K_1, C)$
$dist(K_2, C)$	$dist(K_2, C)$
$dist(K_3, C)$	$dist(K_3, C)$
$dist(\mathbf{K_4}, C)$	$dist(K_3, C)$
Min($dist(K_2, T_1)$, $dist(K_2, T_2)$, $dist(K_2, \mathbf{T_3})$)	Min($dist(K_2, T_1)$, $dist(K_2, T_2)$)
Min($dist(K_3, T_1)$, $dist(K_3, T_2)$, $dist(K_3, \mathbf{T_3})$)	Min($dist(K_3, T_1)$, $dist(K_3, T_2)$)
Min($dist(\mathbf{K_4}, T_1)$, $dist(\mathbf{K_4}, T_2)$, $dist(\mathbf{K_4}, \mathbf{T_3})$)	Min($dist(K_3, T_1)$, $dist(K_3, T_2)$)
Min($ang(K_2, K_1, T_1)$, $ang(K_2, K_1, T_2)$, $ang(K_2, K_1, \mathbf{T_3})$)	Min($ang(K_2, K_1, T_1)$, $ang(K_2, K_1, T_2)$)
Min($ang(K_3, K_1, T_1)$, $ang(K_3, K_1, T_2)$, $ang(K_3, K_1, \mathbf{T_3})$)	Min($ang(K_3, K_1, T_1)$, $ang(K_3, K_1, T_2)$)
Min($ang(\mathbf{K_4}, K_1, T_1)$, $ang(\mathbf{K_4}, K_1, T_2)$, $ang(\mathbf{K_4}, K_1, \mathbf{T_3})$)	Min($ang(K_3, K_1, T_1)$, $ang(K_3, K_1, T_2)$)

Table 1: This table describes part of the ρ transform from states in 3 vs. 2 keepaway to states in 4 vs. 3 keepaway. We denote the distance between a and b as $dist(a, b)$ and the angle formed between players a, c with a vertex at b as $ang(a, b, c)$. Relevant points are the center of the field C, keepers K_1-K_4, and takers T_1-T_3. Keepers and takers are ordered in increasing distance from the ball and state values not present in 3 vs. 2 are in bold.

# of 3 vs. 2 episodes	Ave. 4 vs. 3 time (hours)	Ave. total time (hours)
0	15.26	15.26
1	12.29	12.29
10	10.06	10.08
50	4.83	4.93
100	4.02	**4.22**
250	3.77	4.3
500	3.99	5.05
1000	3.72	5.85
3000	2.42	11.04
9000	1.38	50.49
18000	**1.24**	98.01

Table 2: Results from learning keepaway with different amounts of 3 vs. 2 training time indicate that behavior transfer can reduce training time. The minimum times for learning 4 vs. 3 and for total time are in bold.

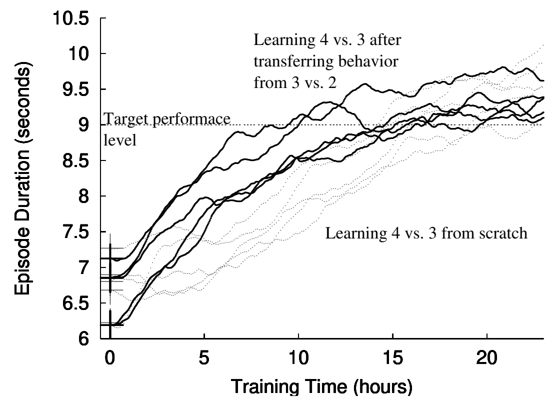

Figure 4: The learning curves for five representative keepers in the 4 vs. 3 keepaway domain when learning from scratch (dotted lines) have similar initial hold times when compared to five representative learning curves generated by transferring behavior from the 3 vs. 2 task (solid lines). The learners which have benefited from behavior transfer are able to more quickly learn the 4 vs. 3 task.

set of experiments. In both cases we found that the learning time was *greater* than learning from scratch. Haphazardly initializing CMAC weights may hurt learners, but systematically setting them through behavior transfer is beneficial.

We would like to be able to determine the optimal amount of time needed to train on an easier task to speed up a more difficult task. While it is not critical when considering the 4 vs. 3 task because many choices produce near optimal results, finding these values becomes increasingly difficult and increasingly important as we scale up to larger tasks, such as 5 vs. 4 keepaway. Determining these training thresholds for tasks in different domains is currently an open problem and will be the subject of future research.

Interestingly, when the CMACs' weights are loaded into the keepers in 4 vs. 3, the initial hold times of the keepers do not differ much from the keepers with uninitialized CMACs, as shown qualitatively in Figure 4. However, the information contained in the CMACs' weights prime the 4 vs. 3 keepers to more quickly learn their task. As the figure suggests, the 4 vs. 3 keepers which have loaded weights from 3 vs. 2 players learn at a faster rate than those 4 vs. 3 players that are training from scratch. This outcome suggests that the learned behavior is able to speed up the rate of reinforcement learning on the novel domain even though the knowledge we transfer is of limited initial value.

The final step of ρ where we put the average learned weight into all uninitialized weights is more beneficial at lower numbers of episodes; it gives initial values to weights in the state/action space that have never been visited. Over time more of the state space is explored and thus the averaging step yields less benefit. While it may seem surprising that a single episode of 3 vs. 2 can significantly effect the total training times, we can see that the averaging effect has a large impact on the total learning time, even for a small number of episodes. Note also that the SARSA update is run on average 23 times times during the initial 3 vs. 2 episode, where each update affects 416 CMAC weights, suggesting something useful can be learned.

To test the sensitivity of the ρ function, we tried modifying it so that instead of copying the weights for the state variables for K_3 into the new 4 vs. 3 K_4 (see Table 1), we instead copy the K_2 state variable to this location. Now $\pi_{4vs3,initial}$ will evaluate the state variables for the closest and furthest keeper teammates to the same value instead of the two furthest teammates. Similarly, instead of copying weights corresponding to T_2 into the T_3 location, we copy weights from T_1. Training on 1,000 3 vs. 2 episodes and using $\rho_{modified}$ to initialize weights in 4 vs. 3, the total training time increased. Although this $\rho_{modified}$ outperformed training from scratch, the total training time is 10%-20% longer compared to using ρ. Choosing non-optimal mappings between actions and states when constructing ρ seems to have a detrimental, but not necessarily disastrous, effect on the training time.

# of 3 vs. 2 episodes	# of 4 vs. 3 episodes	Ave. 5 vs. 4 time (hours)	Ave. total time (hours)
0	0	24.32	24.32
0	1000	17.06	18.38
500	500	10.19	11.75

Table 3: Results from learning 5 vs. 4 keepaway with behavior transfer demonstrate that our method scales to more difficult tasks.

Initial results in Table 3 show that behavior transfer scales to the 5 vs. 4 keepaway task as well. The 5 vs. 4 task is harder than the 4 vs. 3 task for the same reasons that were detailed in Section 4.2 when discussing why 4 vs. 3 is more difficult than 3 vs. 2. In 5 vs. 4 we say that the task has been learned when the 5 keepers are able to hold the ball for an average of 9.0 seconds over 1,000 episodes. If behavior transfer from a learned 4 vs. 3 value function is used, the 5 vs. 4 training time can be reduced by roughly 30% and the total time is reduced by 24% when compared to learning from scratch. An additional refinement is to use a two-step application of behavior transfer so that 3 vs. 2 runs first, this learned value function is used as the initial value function in 4 vs. 3, and after training the final 4 vs. 3 value function is used as the initial value function for 5 vs. 4. Using this procedure we find that the time to learn 5 vs. 4 is reduced by roughly 58% and the total training time is reduced by 52% when compared to learning from scratch. A t-test confirms that the differences between 5 vs. 4 hold times are all statistically significant ($p < 1.17 * 10^{-4}$). We anticipate that behavior transfer will further reduce the total training time necessary to learn 5 vs. 4 as we tune the number 3 vs. 2 and 4 vs. 3 episodes.

7. RELATED WORK

The concept of seeding a learned behavior with some initial simple behavior is not new. There have been approaches to simplifying reinforcement learning by manipulating the transition function, the agent's initial state, and/or the reward function. Directed training [17] is a technique to speed up learning whereby a human is allowed to change the task by modifying the transition function T. Using this method a human supervisor can gradually increase the difficulty of a task while using the same policy as the initial control for the learner. For instance, balancing a pole may be made harder for the learner by decreasing the mass or length of the pole. The learner will adapt to the new task faster using a policy trained on a related task than if learning from scratch.

Learning from easy missions [4] allows a human to change the start state of the learner, $s_{initial}$, making the task incrementally harder. Starting the learner near the exit of a maze and gradually allowing the learner to start further and further from the goal is an example of this. This kind of direction allows the learner to spend less total time learning to perform the final task.

The "transfer of learning" approach [18] applies specifically to temporally sequential subtasks. Using this compositional learning a large task may be broken down into subtasks which are easier to learn and have distinct beginning and termination conditions. However the subtasks must all be very similar in that they have the same state spaces, action spaces, and environment dynamics, although the reward function R may differ.

Another successful idea, reward shaping [6, 11], also contrasts with behavior transfer. In shaping, learners are given an artificial problem which will allow the learner to train faster than on the actual problem which has different environmental rewards, R. Behavior transfer differs in intent in that we aim to transfer behaviors from existing, relevant tasks which can have different state and action spaces rather than creating artificial problems which are easier for the agent to learn. In the RoboCup soccer domain all the different keepaway tasks are directly applicable to the full task of simulated soccer; in keepaway we are always training the learner on useful tasks rather than just simpler variations of a real task. Using behavior transfer, learners are able to take previously learned behaviors from related tasks and apply that behavior to harder tasks which can have different state and action spaces.

While these four methods do allow the learner to spend less total time training, they rely on a human modifying the transition function, the initial start state, or the reward function to create artificial problems to train on. We contrast this with behavior transfer where we allow the state and/or action spaces to change. This added flexibility permits behavior transfer to be applied to a wider range of domains and tasks than the other aforementioned methods. Furthermore, behavior transfer does not preclude the modification of the transition function, the start state, or the reward function and can therefore be combined with the other methods if desired.

In some problems where subtasks are clearly defined by features, the subtasks can be automatically identified [7] and leveraged to increase learning rates. This method is not directly applicable to tasks which do not have features that clearly define subtasks. Furthermore, if the shape of the various regions in the value function are too complex, i.e. the smoothness assumption is violated too often, the algorithm to automatically detect subtasks will fail.

Learned subroutines have been successfully transfered in a hierarchical reinforcement learning framework [2]. By analyzing two tasks, subroutines may be identified which can be directly reused in a second task that has a slightly modified state space. The learning rate for the second task can be substantially increased by duplicating the local sub-policy. This work can be thought of as another example for which ρ has been successfully constructed, but in a very different way.

For tasks which can be framed in a relational framework [8], there is research [12] which suggests ways of speeding up learning between two relational reinforcement learning tasks. However, there is no obvious way to apply this technique to reinforcement learning problems that are not relational.

Imitation is another technique which may transfer knowledge from one learner to another [14]. However, there is the assumption that "the mentor and observer have similar abilities" and thus may not be directly applicable when the number of dimensions of the state space changes or the agents have a qualitatively different action set. The goal of our research different: behavior transfer leverages knowledge between tasks to speed up learning whereas imitation focuses on transferring knowledge from one agent to another.

Other research [9] has shown that it is possible to learn policies for large-scale planning tasks that generalize across different tasks in the same domain. Using this method researchers are able to speed up learning in different tasks without explicitly transferring any knowledge, as the policy is defined for the planning domain rather than a specific task.

Another approach [10] uses linear programming to determine value functions for classes of similar agents. Using the assumption that T and R are similar among all agents of a class, class-based value subfunctions are used by agents in a new world that has a different number of objects (and thus different S and A). Although no learning is performed in the new world, the previously learned value functions may still perform better than a baseline handcoded strategy. However, as the authors themselves state, the technique will not perform well in heterogeneous environments or domains with "strong and constant interactions between many objects (e.g. RoboCup)." Our work is further differentiated as we continue learning in the second domain after performing ρ. While the initial performance in the new domain may be increased after loading learned value functions compared to learning from scratch, we have found that a main benefit is an increased learning rate.

8. CONCLUSIONS

We have introduced the behavior transfer method of speeding up reinforcement learning and given empirical evidence for its usefulness. We have trained learners using TD reinforcement learning in related tasks with different state and action spaces and shown that not only is the time to learn the final task reduced, but that the total training time is reduced using behavior transfer when compared to learning the final task from scratch.

Acknowledgments

We would like to thank Gregory Kuhlmann for his help with the experiments described in this paper as well as Yaxin Liu, Nick Jong, Raymond Mooney, and David Pardoe for helpful comments and suggestions. This research was supported in part by NSF CAREER award IIS-0237699, DARPA grant HR0011-04-1-0035, and the UT-Austin MCD Fellowship.

9. REFERENCES

[1] J. S. Albus. *Brains, Behavior, and Robotics*. Byte Books, Peterborough, NH, 1981.

[2] D. Andre and S. J. Russell. State abstraction for programmable reinforcement learning agents. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 119–125, 2002.

[3] D. Andre and A. Teller. Evolving team Darwin United. In M. Asada and H. Kitano, editors, *RoboCup-98: Robot Soccer World Cup II*. Springer Verlag, Berlin, 1999.

[4] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based behavior acquisition for a shooting robot by using a reinforcement learning. In *Proc. of IAPR/IEEE Workshop on Visual Behaviors-1994*, pages 112–118, 1994.

[5] M. Colombetti and M. Dorigo. Robot Shaping: Developing Situated Agents through Learning. Technical Report TR-92-040, International Computer Science Institute, Berkeley, CA, 1993.

[6] C. Drummond. Accelerating reinforcement learning by composing solutions of automatically identified subtasks. *Journal of Artificial Intelligence Research*, 16:59–104, 2002.

[7] S. Dzeroski, L. D. Raedt, and K. Driessens. Relational reinforcement learning. *Machine Learning*, 43:7–52, 2001.

[8] A. Fern, S. Yoon, and R. Givan. Approximate policy iteration with a policy language bias. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[9] C. Guestrin, D. Koller, C. Gearhart, and N. Kanodia. Generalizing plans to new environments in relational mdps. In *International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, August 2003.

[10] M. J. Mataric. Reward functions for accelerated learning. In *International Conference on Machine Learning*, pages 181–189, 1994.

[11] E. F. Morales. Scaling up reinforcement learning with a relational representation. In *Proc. of the Workshop on Adaptability in Multi-agent Systems*, January 2003.

[12] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. 16th International Conf. on Machine Learning*, 1999.

[13] B. Price and C. Boutilier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629, 2003.

[14] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[15] M. Riedmiller, A. Merke, D. Meier, A. Hoffman, A. Sinner, O. Thate, and R. Ehrmann. Karlsruhe brainstormers—a reinforcement learning approach to robotic soccer. In P. Stone, T. Balch, and G. Kraetszchmar, editors, *RoboCup-2000: Robot Soccer World Cup IV*. Springer Verlag, Berlin, 2001.

[16] O. Selfridge, R. S. Sutton, and A. G. Barto. Training and tracking in robotics. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 670–672, 1985.

[17] S. P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8:323–339, 1992.

[18] P. Stone, P. Riley, and M. Veloso. The CMUnited-99 champion simulator team. In M. Veloso, E. Pagello, and H. Kitano, editors, *RoboCup-99: Robot Soccer World Cup III*, pages 35–48. Springer, Berlin, 2000.

[19] P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup-soccer keepaway. *Adaptive Behavior*, 2005. To appear.

[20] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[21] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

10. REFERENCES

[1] J. S. Albus. *Brains, Behavior, and Robotics*. Byte Books, Peterborough, NH, 1981.

[2] D. Andre and S. J. Russell. State abstraction for programmable reinforcement learning agents. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 119–125, 2002.

[3] D. Andre and A. Teller. Evolving team Darwin United. In M. Asada and H. Kitano, editors, *RoboCup-98: Robot Soccer World Cup II*. Springer Verlag, Berlin, 1999.

[4] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based behavior acquisition for a shooting robot by using a reinforcement learning. In *Proc. of IAPR/IEEE Workshop on Visual Behaviors-1994*, pages 112–118, 1994.

[5] R. Boer and J. Kok. The Incremental Development of a Synthetic Multi-agent System: The UvA Trilearn 2001 Robotic Soccer Simulation Team. Master's thesis, University of Amsterdam, The Netherlands, February 2002.

[6] M. Colombetti and M. Dorigo. Robot Shaping: Developing Situated Agents through Learning. Technical Report TR-92-040, International Computer Science Institute, Berkeley, CA, 1993.

[7] C. Drummond. Accelerating reinforcement learning by composing solutions of automatically identified subtasks. *Journal of Artificial Intelligence Research*, 16:59–104, 2002.

[8] S. Dzeroski, L. D. Raedt, and K. Driessens. Relational reinforcement learning. *Machine Learning*, 43:7–52, 2001.

[9] A. Fern, S. Yoon, and R. Givan. Approximate policy iteration with a policy language bias. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[10] C. Guestrin, D. Koller, C. Gearhart, and N. Kanodia. Generalizing plans to new environments in relational mdps. In *International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, August 2003.

[11] M. J. Mataric. Reward functions for accelerated learning. In *International Conference on Machine Learning*, pages 181–189, 1994.

[12] E. F. Morales. Scaling up reinforcement learning with a relational representation. In *Proc. of the Workshop on Adaptability in Multi-agent Systems*, January 2003.

[13] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. 16th International Conf. on Machine Learning*, 1999.

[14] B. Price and C. Boutilier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629, 2003.

[15] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[16] M. Riedmiller, A. Merke, D. Meier, A. Hoffman, A. Sinner, O. Thate, and R. Ehrmann. Karlsruhe brainstormers—a reinforcement learning approach to robotic soccer. In P. Stone, T. Balch, and G. Kraetszchmar, editors, *RoboCup-2000: Robot Soccer World Cup IV*. Springer Verlag, Berlin, 2001.

[17] O. Selfridge, R. S. Sutton, and A. G. Barto. Training and tracking in robotics. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 670–672, 1985.

[18] S. P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8:323–339, 1992.

[19] P. Stone, G. Kuhlmann, M. Taylor, and Y. Liu. Keepaway Soccer: From Machine Learning Testbed to Benchmark. In *Proceedings of RoboCup International Symposium*, 2005. To appear.

[20] P. Stone, P. Riley, and M. Veloso. The CMUnited-99 champion simulator team. In M. Veloso, E. Pagello, and H. Kitano, editors, *RoboCup-99: Robot Soccer World Cup III*, pages 35–48. Springer, Berlin, 2000.

[21] P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup-soccer keepaway. *Adaptive Behavior*, 2005. To appear.

[22] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[23] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.