Learning Complementary Multiagent Behaviors: A Case Study

Shivaram Kalyanakrishnan and Peter Stone

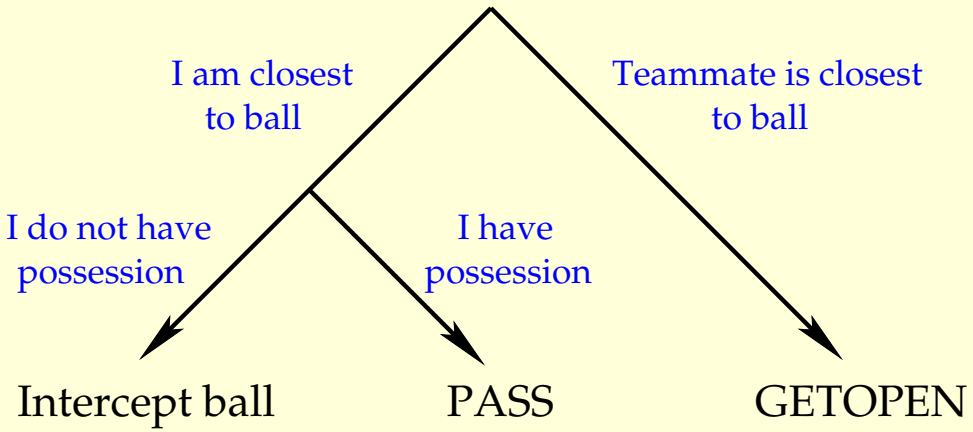
The University of Texas at Austin

May 2009

Motivation: Keepaway Soccer

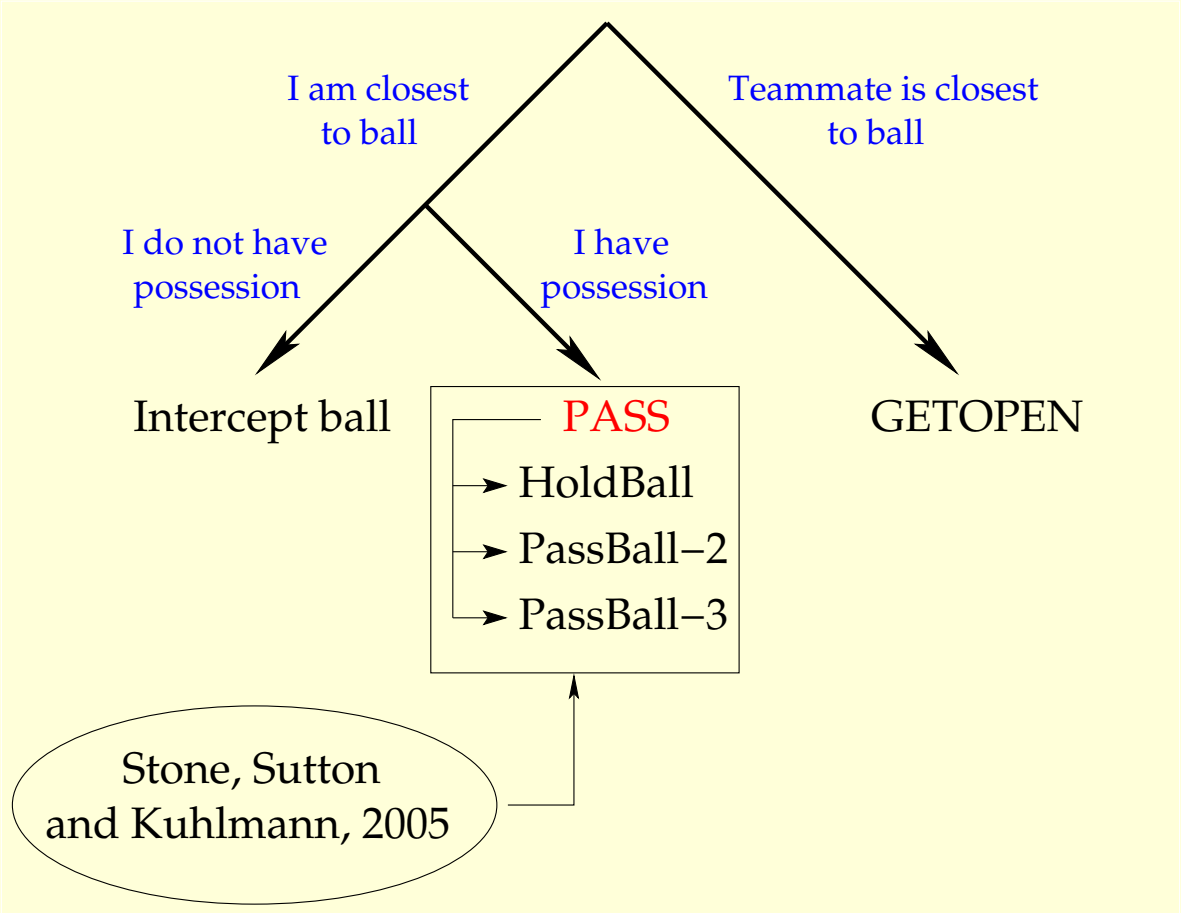
- 3 “keepers”, 2 “takers”.
- Episode ends when takers get possession or ball goes outside field.
- Keepers to maximize episodic **hold time**.
- Noisy sensor information.
- Stochastic, high-level actions.
- Multiagency.
- Real-time processing.

Policy Followed by Each Keeper



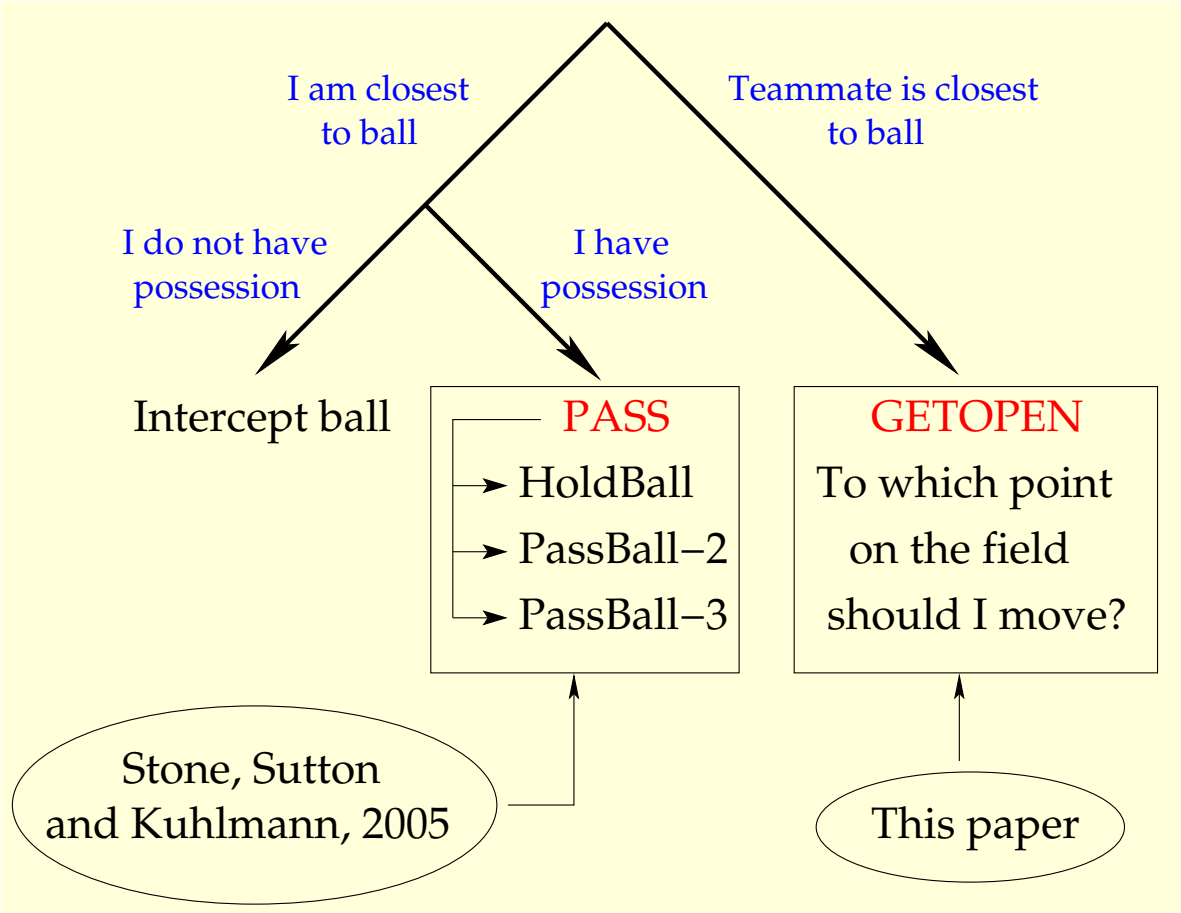
- Takers follow fixed policy of intercepting ball.

Policy Followed by Each Keeper



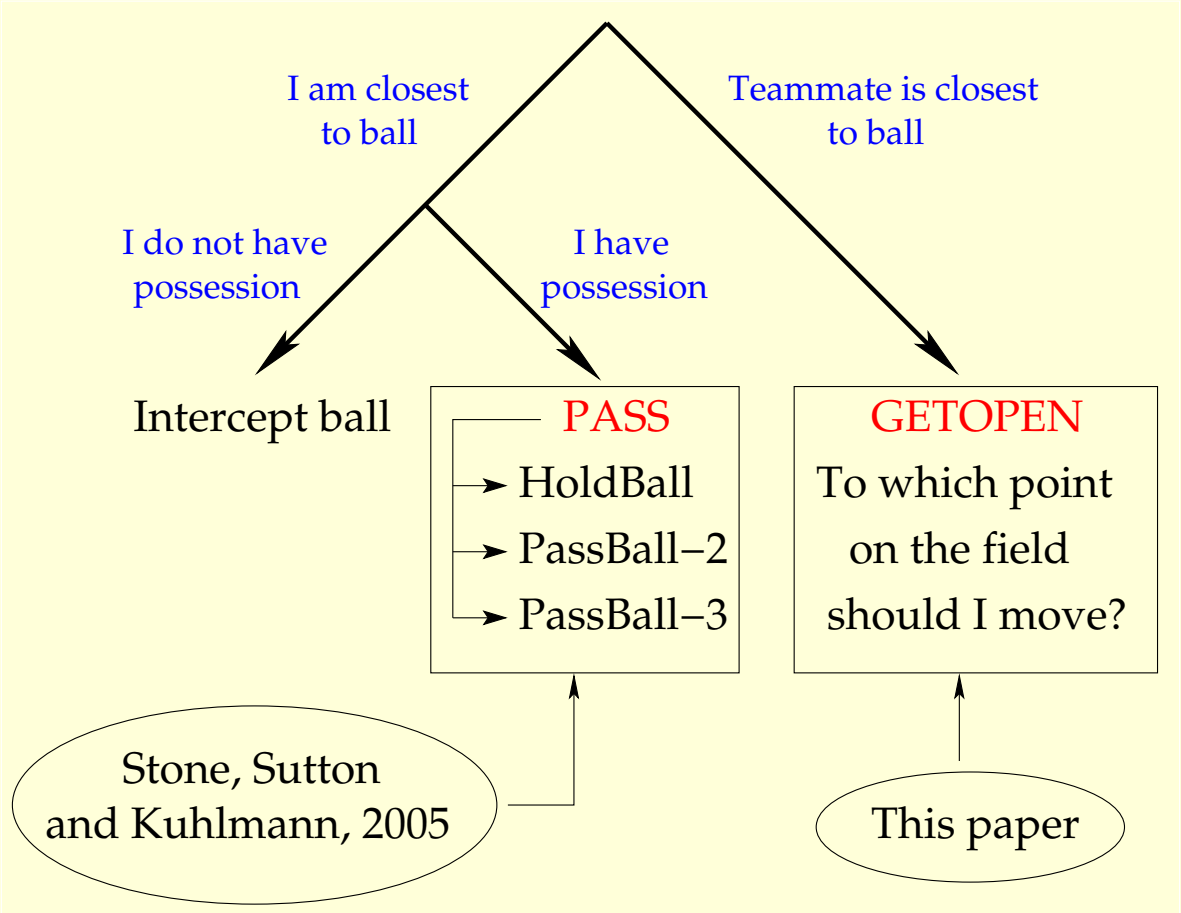
- Takers follow fixed policy of intercepting ball.

Policy Followed by Each Keeper



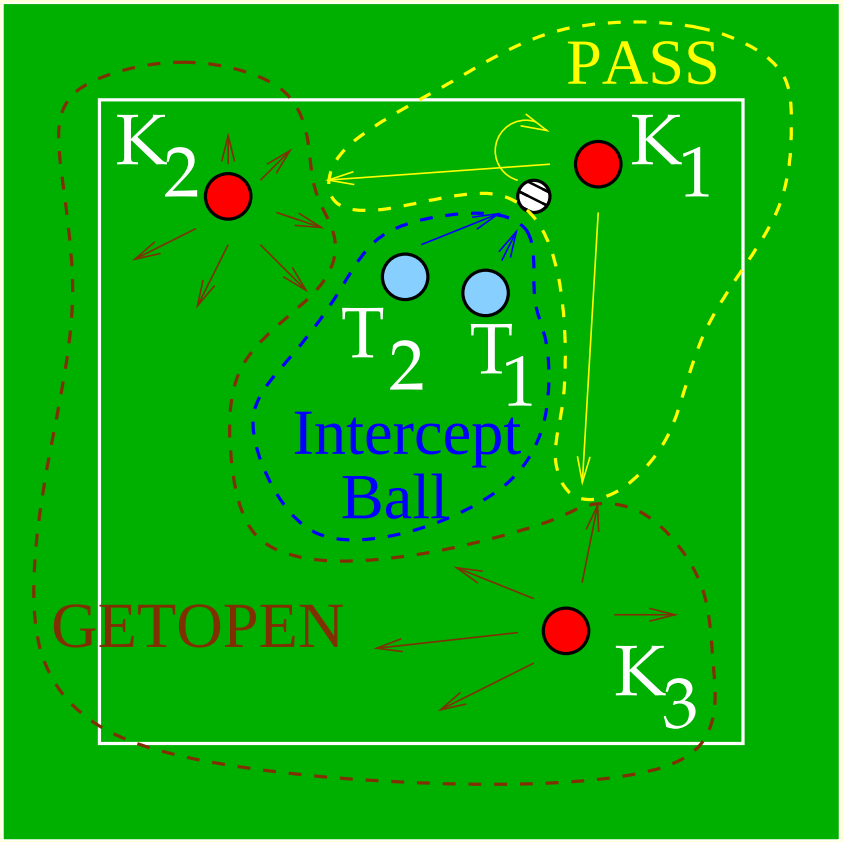
- Takers follow fixed policy of intercepting ball.

Policy Followed by Each Keeper



- Takers follow fixed policy of intercepting ball.

PASS and GETOPEN: Coupled Behaviors



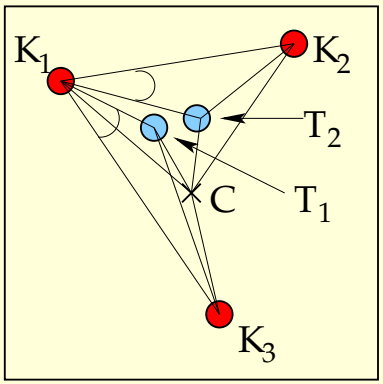
- PASS and GETOPEN fit the category of "distinct populations with coupled fitness landscapes" (Rosin and Belew, 1995).
- *Can we learn* GETOPEN *and* PASS+GETOPEN*?*

Talk Overview

- Motivation
- PASS and GETOPEN: Problem formulation.
- Learning PASS, GETOPEN, and PASS+GETOPEN
- Results
- Related Work
- Conclusion

PASS

- State Variables

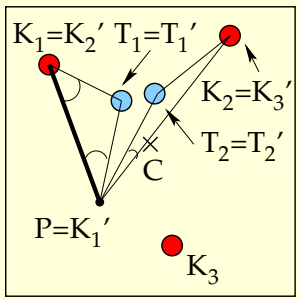


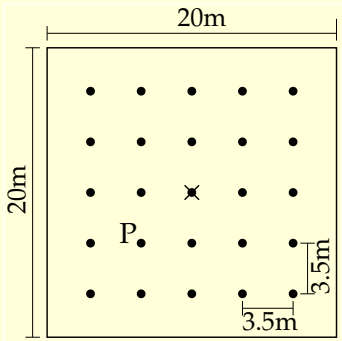
$dist(K_1, C)$ $dist(K_1, K_2)$ $\min_{j \in 1,2} dist(K_2, T_j)$
$dist(K_2, C)$ $dist(K_1, K_3)$ $\min_{j \in 1,2} ang(K_2, K_1, T_j)$
$dist(K_3, C)$ $dist(K_1, T_1)$ $\min_{j \in 1,2} dist(K_3, T_j)$
$dist(T_1, C)$ $dist(K_2, T_2)$ $\min_{j \in 1,2} ang(K_3, K_1, T_j)$
$dist(T_2, C)$

- Actions: {HoldBall, PassBall-2, PassBall-3}.
- To learn policy $\pi : \mathbb{R}^{13} \rightarrow$ {HoldBall, PassBall-2, PassBall-3}.
- PASS policies: PASS:RANDOM, PASS:HAND-CODED, PASS:LEARNED.

GETOPEN

- State Variables

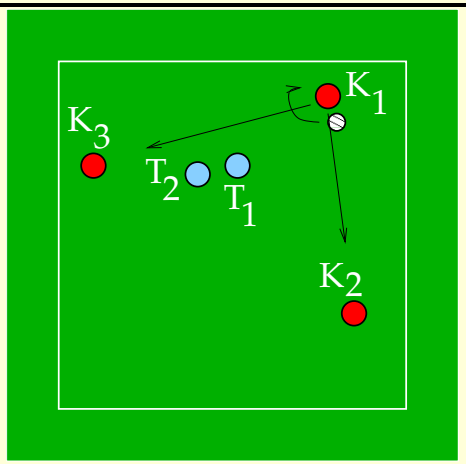
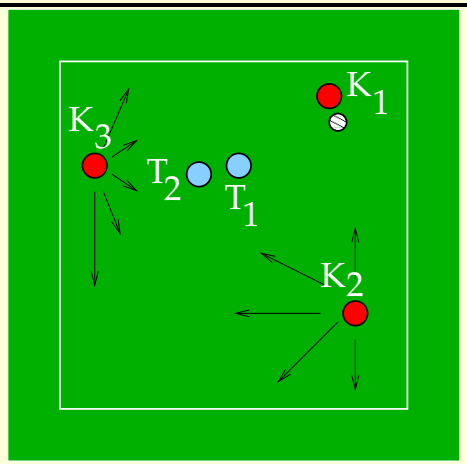




$dist(K_1', K_2')$	$\min_{j \in 1,2} dist(K_2', T_j')$	$dist(K_1, K_1')$
$dist(K_1', K_3')$	$\min_{j \in 1,2} ang(K_2', K_1', T_j')$	$\min_{j \in 1,2 ang(K_1', K_1, T_j)}$
$dist(K_1', T_1')$	$\min_{j \in 1,2} dist(K_3', T_j')$	
$dist(K_2', T_2')$	$\min_{j \in 1,2} ang(K_3', K_1', T_j')$	

- Action: Move to argmax_P *GetOpenValue*(P).
- To learn *GetOpenValue* : $\mathbb{R}^{10} \rightarrow \mathbb{R}$.
- GETOPEN policies: GETOPEN:RANDOM, GETOPEN:HAND-CODED, GETOPEN:LEARNED.

PASS versus GETOPEN

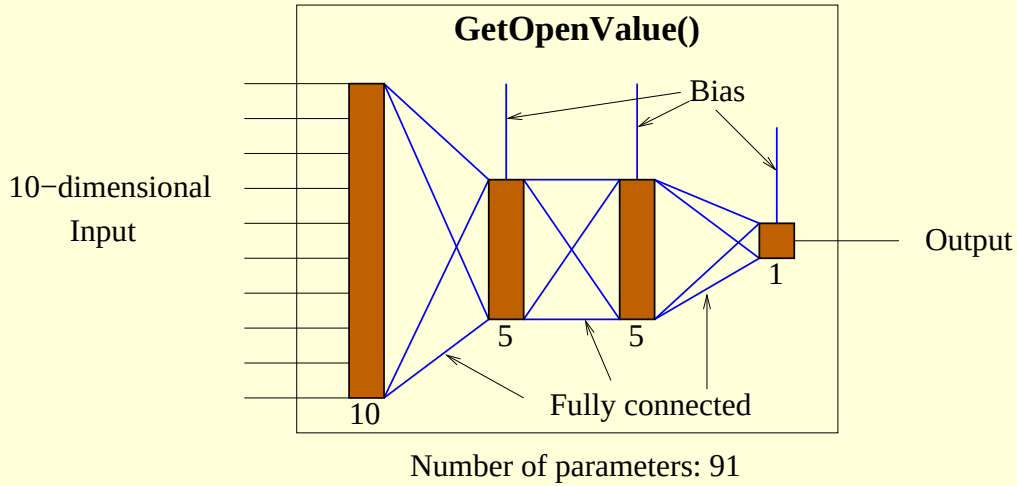
PASS	GETOPEN
	
Executed at a time by one keeper at most, when it has ball possession.	Executed every cycle by two keepers.
3 actions.	25 actions for each keeper.
Objective function can be decomposed into credit for individual actions.	Credit must be given to sequence of joint actions.
Learning methods for PASS and GETOPEN have to cope with non-stationarity if learning PASS+GETOPEN.	

Learning PASS (Stone *et al.*, 2005)

- ϵ-greedy policy ($\epsilon = 0.01$).
- Each keeper makes Sarsa updates every time it take an action or an episode ends:
$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma Q(s',a') - Q(s,a)).$$
- CMAC function approximation of Q, with one-dimensional tilings.
- $\alpha = 0.125$, $\gamma = 1.0$

Learning GETOPEN

- Parameterized representation of solution: 2-layer neural network with sigmoid units.



Number of parameters: 91

- Cross-entropy method for policy search.
  - Generating distribution: Gaussian.
  - Population size: 20.
  - Selection fraction: 0.25.
- Each policy evaluated over 125 episodes of Keepaway and averaged.

Learning PASS+GETOPEN

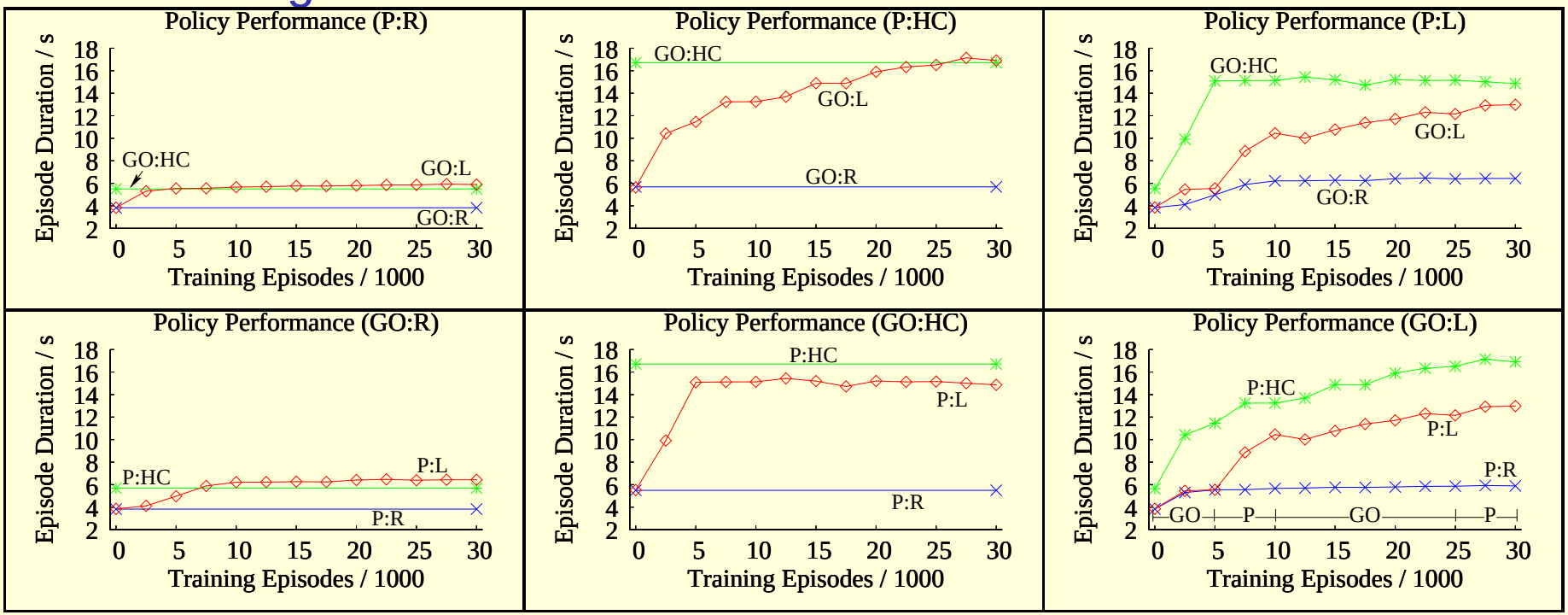
- Interleaved learning: fix PASS policy, learn GETOPEN; fix GETOPEN policy, learn PASS; iterate.

Algorithm 1 Learning PASS+GETOPEN

Output: Policies π_{PASS} and π_{GETOPEN}.
$\pi_{\text{PASS}} \leftarrow$ PASS:RANDOM.
$\pi_{\text{GETOPEN}} \leftarrow$ GETOPEN:RANDOM.
repeat
  $\pi_{\text{GETOPEN}} \leftarrow \textit{learnGetOpen}(\pi_{\text{PASS}}, \pi_{\text{GETOPEN}})$.
  $\pi_{\text{PASS}} \leftarrow \textit{learnPass}(\pi_{\text{PASS}}, \pi_{\text{GETOPEN}})$.
until convergence
Return $\pi_{\text{PASS}}, \pi_{\text{GETOPEN}}$.

- Keepers learn PASS autonomously, but share a common GETOPEN policy.
- In implementation, we allot different numbers of episodes for PASS and GETOPEN.

Results: Learning Performance



- Averages of 20+ independent runs, static evaluation.
- P:HC-GO:L $\approx$ P:HC-GO:HC.
- P:HC-GO:L $>$ P:L-GO:HC.
- P:L-GO:R $>$ P:HC-GO:R.
- P:L-GO:L falls short of P:L-GO:HC, P:HC-GO:L, P:HC-GO:HC.

Results: Specialization of Learned Policies

PASS:LEARNED			
Train	Test		
	GO:R	GO:HC	GO:L
GO:R	**6.37** $\pm$ 0.05	11.73 $\pm$ 0.25	10.54 $\pm$ 0.26
GO:HC	6.34 $\pm$ 0.06^{-}	**15.27** $\pm$ 0.26	12.25 $\pm$ 0.32
GO:L	5.96 $\pm$ 0.07	13.39 $\pm$ 0.35	**13.08** $\pm$ 0.26 (s)
			12.32 $\pm$ 0.32 (d)$^{-}$

GETOPEN:LEARNED			
Train	Test		
	P:R	P:HC	P:L
P:R	**5.89** $\pm$ 0.05	10.40 $\pm$ 0.39	11.15 $\pm$ 0.43
P:HC	5.48 $\pm$ 0.04	**16.89** $\pm$ 0.39	12.99 $\pm$ 0.43^{-}
P:L	5.57 $\pm$ 0.06	11.78 $\pm$ 0.56	**13.08** $\pm$ 0.26 (s)
			12.32 $\pm$ 0.32 (d)$^{-}$

- $(i,j)^{th}$ entry shows performance (and one standard error) of learned policy trained with counterpart i and tested with counterpart j.
- Diagonal entries highest (some not statistically significant).

Results: Videos

	GO:R	GO:HC	GO:L
P:R			
P:HC			
P:L			

Related Work

- Multiple learning algorithms: Stone's "layered learning" architecture (2000) uses neural nets for ball interception, decision trees for evaluating passes, TPOT-RL for temporal difference learning.
- Simultaneous co-evolution: Rosin and Belew (1995) apply genetic evolution in a *competitive* setting on games such as tic-tac-toe and nim. Haynes *et al.* consider cooperative co-evolution in simple predator-prey domain.
- Concurrent and team learning: Panait and Luke's survey (2005).
- Keepaway: Metzen *et al.* (2008) propose "EANT" evolution, Taylor *et al.* (2007) implement behavior transfer. Iscen and Erogul (2008) learn taker behavior.
- Robot soccer: Riedmiller and Gabel (2007) apply model-based reinforcement learning for developing attacker behavior.

Conclusion

- We demonstrate on a significantly complex task the effectiveness of applying qualitatively different learning methods to different parts of the task.
- Learning GETOPEN is at least as rewarding as learning PASS.
- We show the feasibility of learning PASS+GETOPEN, although its performance can be improved.
- We show that tightly-coupled behaviors are learned.
- This work extends the scope of multiagent research in the Keepaway benchmark problem.
- Several avenues of future work arise: replicating research carried out with PASS on GETOPEN, agent communication, etc.