In *RoboCup International Symposium 2010* (Robocup 2010),
Singapore, June 2010.

Learning Powerful Kicks on the Aibo ERS-7: The Quest for a Striker

Matthew Hausknecht and Peter Stone
Department of Computer Science, The University of Texas at Austin
{mhauskn,pstone}@cs.utexas.edu

Abstract. Coordinating complex motion sequences remains a challenging task for robotics. Machine Learning has aided this process, successfully improving motion sequences such as walking and grasping. However, to the best of our knowledge, outside of simulation, learning has never been applied to the task of kicking the ball. We apply machine learning methods to optimize kick power entirely on a real robot. The resulting learned kick is significantly more powerful than the most powerful hand-coded kick of one of the most successful RoboCup four-legged league teams, and is learned in a principled manner which requires very little engineering of the parameter space. Finally, model inversion is applied to the problem of creating a parameterized kick capable of kicking the ball a specified distance.

1 Introduction

Robocup is an international research and education initiative aiming to develop a team of fully autonomous humanoid robots that, by the year 2050, can win against the human-soccer world championship team. In annual Robocup competitions, teams of robots are pitted against each other in a soccer match with both field and rules resembling their human equivalents. For robots, as for humans, the ability to effectively attack the goal and score is crucial. Thus powerful, accurate kicks are highly prized.

Typically, kicks and other motion sequences are painstakingly created and manually tuned. In the last few years however, machine learning has optimized motion sequences associated with the skills of walking quickly [10, 14, 13], receiving passes [12], and capturing the ball [7]. The use of machine learning to optimize a skill has the benefits of removing human bias from the optimization process and, in many cases, reducing the amount of human labor required to create a top notch motion.

We optimize kicking distance on the Sony Aibo ERS-7 using Hill Climbing and Policy Gradient algorithms. The resulting learned kick is compared to the most powerful hand-coded kick of one of Robocup's perennial top competitors, UT Austin Villa [1]. Experiments conducted over multiple robots show the learned kick significantly outperforms UT Austin Villa's most powerful hand-coded kick.

Like scoring, good passing is an important team skill. Unlike scoring, the most powerful kick is not always the best choice for successful passing. While a powerful kick may be able to move the ball to the receiving Aibo, the superfluous

velocity complicates the process of receiving the pass and introduces possible bounce-back concerns. To address these issues, we create a parameterized kick capable of accurately moving the ball a desired distance. As in human soccer, the passing Aibo needs only estimate the distance to the receiver and adjust the power of its kick accordingly.

The remainder of the paper is organized as follows: Section 2 presents the background of learning skills and kicking. Section 3 covers the parameterization of the kick. Section 4 discusses the physical framework in which kicks were learned. Section 5 introduces the algorithms used for learning. Section 6 presents the primary results of optimizing kick distance while Section 7 covers additional results and the variable distance kick. Conclusions and future work are presented in Section 8.

2 Background and Related Work

Related work can be divided into three main categories, namely optimizing RoboCup skills, modeling the effects of Aibo kicks, and using simulation to learn more effective kicks.

Robocup skill learning has optimized the skills of walking quickly, walking with stability, ball grasping, and ball trapping. Quick walking has been optimized by several groups [4, 9–11, 14], with learned gaits routinely outperforming their handcoded counterparts. Among the gait learners, Kohl and Stone [13] compared the performance of hill climbing, amoeba, genetic, and policy gradient algorithms on the task of optimizing Aibo walk speed. Results show that the learned walked significantly outperforms the best hand-coded walks and rivals other learned walks of the time. Subsequently, camera stability was incorporated into the learning process, allowing Saggar et al. [15] to learn a walk which sacrifices a small amount of speed in exchange for superior image stability.

The ball-grasping skill was optimized by Fidelman and Stone [7] using a Layered Learning framework [16]. The final ball grasping behavior shows significant improvements over hand-coded equivalents. Similarly, Kobayashi et al. [12] learn how to trap the ball using Reinforcement Learning [18]. Trapping, as opposed to grasping, is the process of intercepting and capturing a moving ball rather than a stationary one. Results show learning stabilizes near 80% successful trapping accuracy after approximately 350 episodes.

One of the main appeals of using machine learning is the reduction of manual effort. Thus, in each of the skill optimization approaches mentioned above, a physical learning framework was created to minimize the amount of human interference necessary throughout the learning process. In the above examples, humans were only required to start the learning process and to replace the Aibo's battery once drained. The robots were able to autonomously test and evaluate different variations of the skill being learned, keeping track of their own progress. Like previous work, we create a learning framework to minimize the human interference in the learning process. The main novelty in our work is that, to the best of our knowledge, we are the first to use machine learning to optimize

the kicking skill fully outside of simulation, using a much higher dimensional paramaterization than those of the aforementioned skills.

A second class of related work has focussed on modeling the effects of kicks. Chernova and Veloso [5] examined the effects of an arm and a head kick by tracking the angle and displacement of the ball directly after the kick was performed. They subsequently use this knowledge to select the most appropriate kick for a given situation, effectively reducing the time taken for an Aibo to score a goal. Similarly, Ahmadi and Stone [2] investigate kick models for use in the action planning domain. Unlike the parameterized kick models used by Chernova and Veloso, Ahmadi employs instance based models, in which the resulting ball location of each kick is recorded. These more detailed models allow the robot to plan more thoroughly about the effects of a given kick in the presence of obstacles or other robots. Rather than modeling a number of discrete kicks, each approximating a desired effect, we create a parameterized kick capable of moving the ball any reasonable desired kick distance.

Third, the process of learning to kick has been explored primarily inside of the realm of simulation. Zagal and Ruiz-del-Solar [20] used a simulator to investigate reactive ball kicks. Decomposing each kick into four 25-parameter poses (configurations of joint angles for each limb), they successfully learned a falling kick – qualitatively similar to most powerful kicks of the time. Cherubini, Giannone, and Iocchi [6] learned two kicks as a part of a larger, Layered Learning approach focused on learning to attack the goal. Learning started on a simulator and was subsequently continued on the physical robot, where the best simulator policies became the robot's initial policies. Results show progress on the overall task of attacking the goal, but are not reported for the subtask of improving the kicks. Finally, Hester et al. [8] recently applied the RL-DT model-based reinforcement learning algorithm to the task of scoring penalty goals. Learning was performed in simulation (85% success rate) and validated on the real robot (70% success rate). Unlike previous work, our approach has no simulation component. To the best of the author's knowledge, this is the first time a kick has been learned fully on physical robots.

In order to survey state of the art kicks, eleven hand-coded kicking motions were examined from UT Austin Villa's Legged League, a team who has competed in Robocup since 2003, entering the quarterfinals in '04, '05, '07, and the semifinals in '08. Each kick consisted of 4-19 discrete poses (full-body joint angle specifications) and the amount of time to hold each pose. Of the eleven kicks examined, the "Power Kick" [17] was the clearly the most powerful and accurate. This kick grasps the ball between the Aibo's front legs and slams the head down, knocking the ball straight ahead of the Aibo. For the remainder of this work, all learned kicks are benchmarked against the Power Kick.

3 Kick Parameterization

In any learning process, the choice of parameterization can dramatically affect results. Increasing the number of parameters increases the learning time and complexity of the problem. On the other hand, decreasing the number of param-

eters limits the scope of possible learning, in some cases simplifying the problem to the point of triviality. Previously, Kohl used 12 parameters to characterize a walk and Kobayashi used only two in order to learn the ball trapping skill. In simulation Zagal learned kicking with 100 parameters.

We use 66 parameters to characterize a kick – an increase from previous non-simulation learning (Chernova's 54 parameter walk [4] comes closest). Since the Aibo has 18 total joints, a maximum parameterization for a 6 pose kick would use 108 joint parameters ($6 * 18$) and 6 pose timing parameters for a total of 114 parameters. We reduced the maximum parameter space in two ways: first, exploiting the Aibo's bilateral symmetry, the three joints in the front left leg were made to mirror those of the front right; the same was true for the back legs. Second, the robot's two tail joints were considered useless and eliminated. These changes reduced the number of parameters per pose from 18 to 10 – for a total of 66 parameters. By only exploiting the inherent symmetry of the robot, our parameterization retains as much flexibility as possible. Attempts were made to use an even more flexible paramaterization – one not exploiting symmetry – but there were no obvious advantages. Figure 1 displays the 10 parameters used in each pose and the number of possible values each parameter could assume. Integer values were required by the ERS-7 hardware.

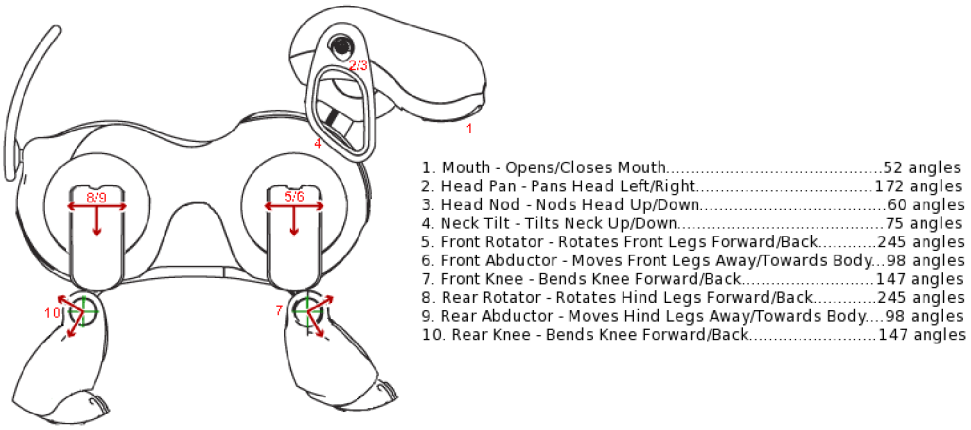

Fig. 1. The ten joints per pose to which learning was applied and the number of possible angles each joint could assume.

As a final note, this parameterization required very little ad hoc engineering or knowledge of the domain. To illustrate this point, consider the 12 parameters used by Kohl to optimize Aibo walk speed: Six of them specify length, height, and shape of the elliptical locus through which the end effector of the foot moves. Two more specify the height of the front and rear body while another determines the fraction of time each foot remains on the ground. Clearly learning has been abstracted to a higher level than specifying raw joint angles. Learning in such a manner often proves more time efficient but requires extensive engineering of the parameter space. In this case it would be necessary, at the minimum, to

engineer functions moving each foot through an adjustable elliptical locus while simultaneously raising and lowering front and rear body height. In contrast, our learning process operates on raw joint angles, requiring no intermediate functionality.

4 Learning to Kick

Figure 2 depicts the inclined ramp constructed to partially automate the kick learning process. Learning consisted of the Aibo kicking the ball up the ramp, receiving feedback on the distance of the kick, and being repositioned for the next kick. The slope of the ramp was adjustable by changing the height of the object the ramp was resting upon. A suitable slope was found through trial an error, guided by the principle that the ramp should be as low as possible so long as the Aibo was unable to kick the ball off the far end.

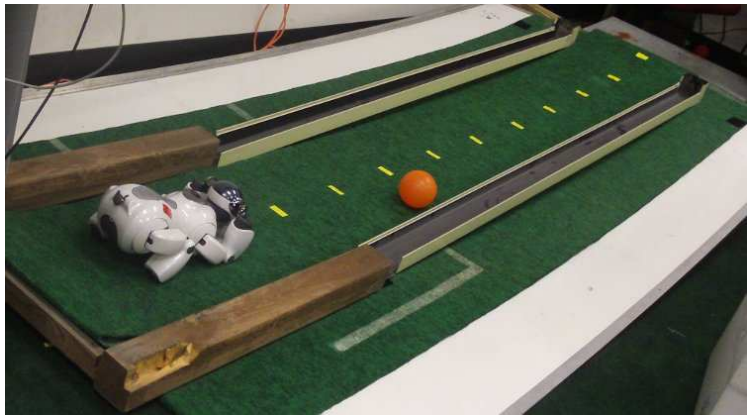

Fig. 2. Inclined ramp for optimizing kick distance.

Despite attempts to achieve full autonomy, a human was required to relay the distance of the kick and reposition the Aibo after each kick. Making this process fully autonomous proved difficult for several reasons: first, the kicking process was violent and unpredictable; It was common for the Aibo to either fall over or fumble the ball backwards, making it hard to recover and reposition the ball for the next kick. Second, because of the carpet covering the ramp, the ball would occasionally get stuck at the apex of its path. This could potentially have been resolved with a steeper incline, but that would have led to shallower kicks and less informative distance feedback. Finally, it would have been problematic for the Aibo to autonomously estimate kick distance primarily because many kicks ended with the Aibo's head (and sometimes body) facing in a different direction than that in which the ball was kicked. Estimating distance for these kicks would require extremely fast visual re-acquisition and distance estimation before the ball reached its apex. Despite the dependence on human assistance, the ramp system was quite efficient, yielding approximately 500 kicks over the

duration of a single battery charge, approximately one hour – an average of one trial every seven or eight seconds.

5 Machine Learning Algorithms

In order to learn the kick within the framework described above, we tested two existing optimization algorithms. Specifically, we used the Hill Climbing and Policy Gradient algorithms exactly as described by Stone and Kohl in their work on quadruped locomotion [13]. The Policy Gradient algorithm was used with a step size η of 2.0 and run for 65 iterations, in each of which 5 new policies were generated and evaluated. Likewise, Hill Climbing used 5 policies per iteration for 27 iterations. The number of iterations used is a result of running each algorithm for the course of a single battery charge. In both algorithms, new policies were generated by incrementing each joint angle by a random number in the range of zero to one-tenth of the full output range of that joint.

The Hill Climbing algorithm had the advantages of simplicity and robustness in the presence of high number of parameters. Hill Climbing starts with an initial policy π consisting of initial values for each of the parameters and generates one or more additional policies $\{R_1, R_2, ..., R_t\}$ in the parameter space around π. Each policy is then evaluated and the best is kept as the initial policy in the next iteration. Because Hill Climbing is little more than a guess and check algorithm, it can safely be used in the presence of high dimensional parameter spaces, making it ideal for both tuning and learning high dimensional kicks.

The Policy Gradient Algorithm was chosen because it had previously proven effective at learning Aibo walking and grabbing. The specific variant of general policy gradient learning techniques [3, 19] seeks to estimate the gradient of the objective function by sampling policies in the vicinity of the initial policy and inferring the partial derivative for each parameter. After the gradient has been estimated, the new policy is shifted in the direction maximizing the objective function.

6 Results

The main result of this work is the creation of a kick significantly more powerful than the Power Kick. Though anecdotal evidence suggests that other Robocup teams may have had kicks stronger than the Power Kick, it is the strongest kick from the UT Austin Villa arsenal - an arsenal that, as evidenced by the team's success, was effective in practice. To achieve this result, Policy Gradient and Hill Climbing algorithms were applied in succession to create the final learned kick. Starting from UT Austin Villa's Power Kick, Policy Gradient was run for 65 iterations (650 kicks) over the course of a single battery charge. In each iteration, ten policies were evaluated with the score for each policy determined by allowing the policy to generate a single kick and timing how long it took for the ball to roll up the ramp and return the Aibo. Figure 3 plots the score of the first policy at each iteration with higher millisecond return times corresponding to more powerful kicks. Additionally, the sum of squared difference for each parameter

(angle) is plotted between the current policy and the initial policy. This measure indicates whether or not the current policy is stuck inside of a local optimum around the initial policy or if it is successfully traversing the parameter space.

As the solid line in Figure 3 indicates, Policy Gradient (PG) effectively explores the parameter space without getting stuck in a local optimum around the initial policy. Generally, while there were occasional spikes in score, as indicated by the dashed line, there is little sustained improvement in overall score. Even though PG did not converge to a single kick, it did identify, in isolation, several promising motion sequences. The single most powerful kick (iteration 47) from the 65 iterations of PG was selected and further refined through 21 iterations of Hill Climbing, in each of which 5 policies were generated and scored by averaging the power of two kicks generated by that policy.

At this point, the Learned Kick was evaluated against UT Austin Villa's Power Kick on flat ground with ten different Aibos and five trials of each kick per Aibo. Figure 4 displays the resting locations of the 50 Power and Learned Kicks. On average, the Learned Kick moved the ball 373.66cm with a standard deviation of 105.51cm compared to the Power Kick which had an average kick distance of 322.4cm and a standard deviation of 130.21cm. This increase proves statistically significant with a two-tailed P value of 0.0330 on an unpaired t-test. Additionally, the learned kick was able to always move the ball at least 200cm and yielded one kick so powerful that it was stopped only after hitting the far wall of the room, a distance of 628cm, or over 20 feet!

Accuracy for both kicks was assessed by recording how many centimeters the kick deviated from the X-axis per meter of kick distance. According to this metric, the Power Kick is slightly, but not statistically significantly, more accurate than the Learned kick, with an average Y-deviation of 15.82cm per meter of kick distance (std. dev. 14.88) compared to the Learned Kick's 19.17cm Y-deviation (std. dev. 16.32). This effect is likely a result of accuracy not being emphasized during the learning process, aside from the slight loss of momentum when an inaccurate kick would hit a wall of the inclined ramp.

Figures 5 and 6 show the poses and action of the learned kick. Videos of both kicks can be found online at http://www.youtube.com/user/aiboKick.

7 Additional Results

The results in Section 6 demonstrate that learning a kick can yield a stronger kick than laborious hand-tuning. In this section, we further examine the effect of the starting kick on the learning process, as well as whether the kick can be modified for variable distances.

7.1 Effect of Starting Kick

Our final kick was based on learning from an already-tuned starting point. In this section we demonstrate that learning can also work from a weak starting point. Figure 7 shows the results of 27 iterations of Hill Climbing (270 kicks) from a motion sequence unable to propel the ball forwards.

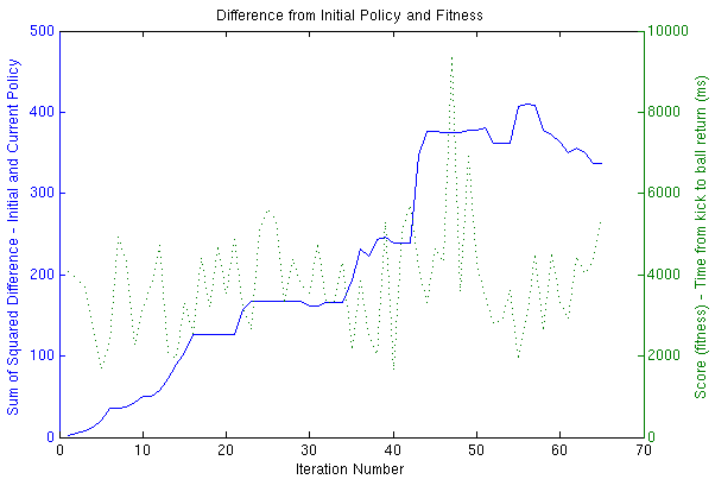

Fig. 3. Policy Gradient learning from Power Kick.

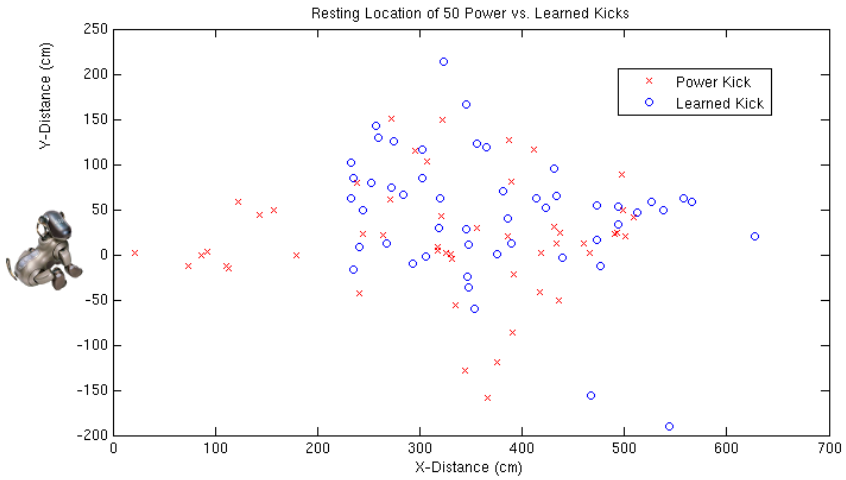

Fig. 4. Aibo kicks from origin in positive X-direction.

Fig. 5. The six poses comprising the final learned kick.

Fig. 6. Learned Kick in motion.

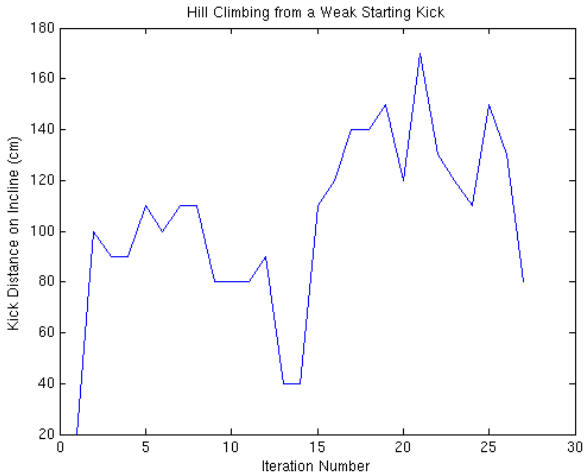

Fig. 7. Hill Climbing from a weak starting kick.

Five policies were evaluted in each of the 27 iterations of hill climbing, with the best becoming the initial policy for the next iteration. Each policy was scored by averaging two kicks generated by that policy. The power of the kick at each iteration was evaluated by noting how far up the 200cm ramp the ball traveled. The initial policy started with a distance score of zero, as it was unable to kick the ball forwards. In just a few iterations a working kick was discovered and generally optimized throughout the rest of the run. After only 21 iterations, the kick was powerful enough to near the end of the 200cm ramp. Admittedly, not any initial policy will lead to a working kick. Fully random initial policies were experimented with and generally had little potential for creating working kicks. In contrast, the starting motion sequence used here had the advantage of keeping the Aibo roughly in a standing position for all 6 poses.

7.2 Variable Distance Kick

While a powerful and accurate kick might be able to reliably reach a target Aibo, we speculate that overpowered passes are harder to receive than passes with a predictable, and more moderate, velocity. Accordingly, we seek to design a kick which gives the passing Aibo control over the power and distance of the hit.

Rather than learning a new kick, or possibly many new kicks – one for each desired distance, we investigate how changes in the timings between poses on an existing kick affect that kick's power. Both UT Austin Villa's Power Kick and the Learned Kick (Section 6) consist of 6 poses with 6 pose transition times. In both, the fourth pose was identified as the "hit" pose – the pose in which the Aibo makes contact with and propels the ball forwards. It was observed that varying the amount of time allotted to this pose greatly affected the power of the kick.

Distances of Power and Learned kicks on flat ground were recorded for number of frames in "hit" pose ranging from 0 to 100, with the Figure 8 showing the

interesting part of each range. Each data point represents the average distance of three kicks.

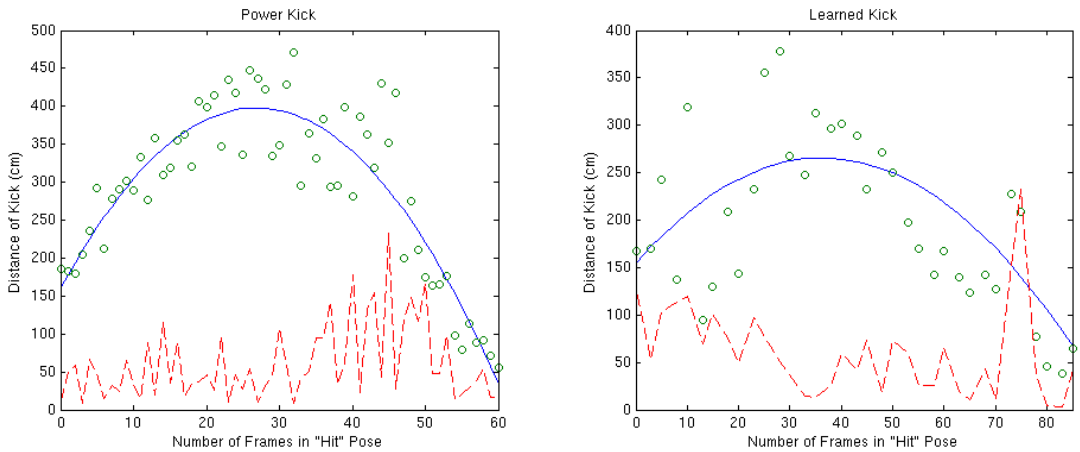

Fig. 8. Quadratics (solid line) fit to kick distance data points (circles) and standard deviations of the 3 kicks comprising each data point (dashed line).

After distance data was collected, a quadratic function was fit to the points. In order to invert the model, it was necessary to solve for the number of frames as a function of the desired kick distance. Table 1 displays the two solutions found for each kick.

	Power Kick	Learned Kick
Quad Eq.	$y = -.3x^2 + 17.7x + 160.1$	$y = -.08x^2 + 6.1x + 154.6$
Soln. 1	$x = -(\sqrt{15} * \sqrt{87437 - 220 * y} - 885)/33$	$x = -(5 * \sqrt{5} * \sqrt{44561 - 168 * y} - 1525)/42$
Soln. 2	$x = (\sqrt{15} * \sqrt{87437 - 220 * y} + 885)/33$	$x = (5 * \sqrt{5} * \sqrt{44561 - 168 * y} + 1525)/42$

Table 1. Quadratic equations fit to kick distance data and their solutions. x is the number of frames in "hit" pose and y is the desired kick distance. Solution 1 corresponds to the left side of the quadratic while Solution 2 fits the right side.

In the case of the Power Kick, the standard deviations shown in Figure 8 indicate that the left side of the quadratic has lower variance and should thus be preferred over the right side whenever possible. As a result, a simple variable distance kick was designed which uses Solution 1, corresponding to the left side of the quadratic, if the desired kick distance is between 160cm and 398cm (the maximum of the quadratic function) and Solution 2, corresponding to the right side of the quadratic, for all other distances. Because the Learned Kick generally showed smaller standard deviations on the right side of the quadratic, Solution 2 was preferred for all distances.

To evaluate the control of both kicks, 20 desired kick distances in the range of 60cm to 400cm were randomly generated. Both kicks were then evaluated based on how close to each desired distance they were able to propel the ball. On average the Power Kick was accurate to within 57.8cm of the requested distance (standard deviation of 41.1cm) and the Learned Kick was accurate to within 45.5cm (standard deviation 39.3cm). To phrase it differently, both kicks were able to place the ball within 1-2 Aibo lengths of their target. It seems likely that the methods applied here to convert a normal kick into a variable distance kick would be equally applicable to nearly any type of kick.

8 Future Work and Conclusions

In this work we demonstrated the learning of a kick more powerful than the strongest hand-coded kick in UT Austin Villa's arsenal. More importantly, this kick was learned in a way that required minimal task-specific engineering of the parameter space. In addition, we created a parameterized kick capable of propelling the ball a requested distance. Source code used for all experiments can be found at http://userweb.cs.utexas.edu/users/AustinVilla/?p=research/aibo_kick.

This paper opens up several interesting directions for future work. First, a more complex inclined ramp could be created to increase the autonomy of the learning process. Additionally, it would have been desirable to benchmark the Learned Kick against the best hand coded kicks of a variety of different teams. Furthermore, accuracy has yet to be incorporated into the learning process. This could potentially be accomplished by recording the distance from the kicking Aibo to either the apex of the kick or the first collision with one of the incline walls. It would be meaningful to evaluate the Variable Distance Kick in the context of the Robocup passing challenge, demonstrating the connection between controllable kicks and successful passes. Finally, it would be interesting to see if a similar learning process could be used for learning to kick on humanoid robots.

9 Acknowledgements

The authors would like to thank Michael Quinlan, Shivaram Kalyanakrishnan, and Todd Hester for useful discussions, as well as the entire UT Austin Villa robot soccer team for development of their code base. Additional thanks to Julian Bishop for help constructing the inclined ramp. This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (CNS-0615104 and IIS-0917122), ONR (N00014-09-1-0658), DARPA (FA8650-08-C-7812), and the Federal Highway Administration (DTFH61-07-H-00030).

References

1. UT Austin Villa Website: http://www.cs.utexas.edu/users/AustinVilla/.

2. M. Ahmadi and P. Stone. Instance-based action models for fast action planning. In U. Visser, F. Ribeiro, T. Ohashi, and F. Dellaert, editors, *RoboCup-2007: Robot Soccer World Cup XI*, volume 5001 of *Lecture Notes in Artificial Intelligence*, pages 1–16. Springer Verlag, Berlin, 2008.
3. J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Intell. Res. (JAIR)*, 15:319–350, 2001.
4. S. Chernova and M. Veloso. An evolutionary approach to gait learning for four-legged robots. In *In Proceedings of IROS'04*, September 2004.
5. S. Chernova and M. Veloso. Learning and using models of kicking motions for legged robots. In *ICRA*, May 2004.
6. A. Cherubini, F. Giannone, and L. Iocchi. Layered learning for a soccer legged robot helped with a 3d simulator. pages 385–392, 2008.
7. P. Fidelman and P. Stone. The chin pinch: A case study in skill learning on a legged robot. In G. Lakemeyer, E. Sklar, D. Sorenti, and T. Takahashi, editors, *RoboCup-2006: Robot Soccer World Cup X*, volume 4434 of *Lecture Notes in Artificial Intelligence*, pages 59–71. Springer Verlag, Berlin, 2007.
8. T. Hester, M. Quinlan, and P. Stone. Generalized model learning for reinforcement learning on a humanoid robot. In *ICRA*, May 2010.
9. G. Hornby, S. Takamura, J. Yokono, O. Hanagata, T. Yamamoto, and M. Fujita. Evolving robust gaits with aibo. In *IEEE International Conference on Robotics and Automation*, pages 3040–3045, 2000.
10. G. S. Hornby, M. Fujita, and S. Takamura. Autonomous evolution of gaits with the sony quadruped robot. In *in Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1297–1304. Morgan Kaufmann, 1999.
11. M. S. Kim and W. Uther. Automatic gait optimisation for quadruped robots. In *Australasian Conference on Robotics and Automation*, Brisbane, December 2003.
12. H. Kobayashi, T. Osaki, E. Williams, A. Ishino, and A. Shinohara. Autonomous learning of ball trapping in the four-legged robot league. pages 86–97, 2007.
13. N. Kohl and P. Stone. Machine learning for fast quadrupedal locomotion. In *The Nineteenth National Conference on Artificial Intelligence*, July 2004.
14. M. J. Quinlan, S. K. Chalup, and R. H. Middleton. Techniques for improving vision and locomotion on the sony aibo robot. In *In Proceedings of the 2003 Australasian Conference on Robotics and Automation*, 2003.
15. M. Saggar, T. D'Silva, N. Kohl, and P. Stone. Autonomous learning of stable quadruped locomotion. In G. Lakemeyer, E. Sklar, D. Sorenti, and T. Takahashi, editors, *RoboCup-2006: Robot Soccer World Cup X*, volume 4434 of *Lecture Notes in Artificial Intelligence*, pages 98–109. Springer Verlag, Berlin, 2007.
16. P. Stone. *Layered Learning in Multiagent Systems: A Winning Approach to Robotic Soccer*. MIT Press, 2000.
17. P. Stone, K. Dresner, P. Fidelman, N. K. Jong, N. Kohl, G. Kuhlmann, M. Sridharan, and D. Stronger. The UT Austin Villa 2004 RoboCup four-legged team: Coming of age. Technical Report UT-AI-TR-04-313, The University of Texas at Austin, Department of Computer Sciences, AI Laboratory, October 2004.
18. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. Mit Pr, May 1998.
19. R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *In Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 1999.
20. J. C. Zagal and J. Ruiz-del-Solar. Learning to kick the ball using Back-to-Reality. In *RoboCup 2004: Robot Soccer World Cup VII*, volume 3276 of *Lecture Notes in Computer Science*, pages 335–346. Springer, 2004.