Proceedings of the Twenty-Sixth International Conference on
Automated Planning and Scheduling (ICAPS 2016)

Online Learning of Robot Soccer Free Kick Plans Using a Bandit Approach

Juan Pablo Mendoza, Reid Simmons and Manuela Veloso
School of Computer Science. Carnegie Mellon University
5000 Forbes Avenue. Pittsburgh, PA 15213

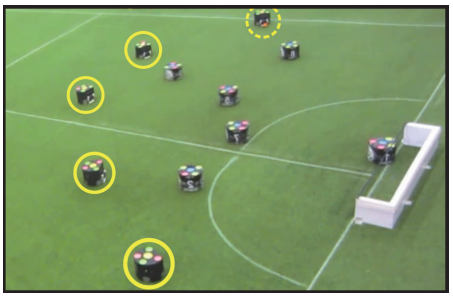
Abstract

This paper presents an online learning approach for teams of autonomous soccer robots to select free kick plans. In robot soccer, free kicks present an opportunity to execute plans with relatively controllable initial conditions. However, the effectiveness of each plan is highly dependent on the adversary, and there are few free kicks during each game, making it necessary to learn online from sparse observations. To achieve learning, we first greatly reduce the planning space by framing the problem as a contextual multi-armed bandit problem, in which the actions are a set of pre-computed plans, and the state is the position of the free kick on the field. During execution, we model the reward function for different free kicks using Gaussian Processes, and perform online learning using the Upper Confidence Bound algorithm. Results from a physics-based simulation reveal that the robots are capable of adapting to various different realistic opponents to maximize their expected reward during free kicks.

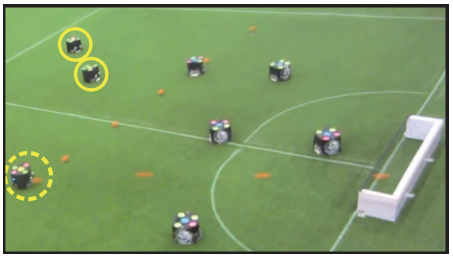
Introduction

Online learning is an appealing and challenging problem in the domain of autonomous robot soccer. Online adaptation is essential to optimize performance against previously unknown opponents with varied strategies. However, the planning space of the team is extremely large, and the robots only have a few minutes of execution to adapt. This paper focuses on online learning for *offensive free kicks* –i.e., free kicks taken by our team from the opponent's half of the field– for which we would want to find weaknesses in the opponent's marking, leading to repeated scoring.

In a game of the RoboCup Small Size League (SSL) [1] soccer, there are around 10 to 20 offensive free kicks per game, making it necessary to adapt from sparse data. To this end, we approach the problem as a multi-armed bandit problem (Gittins, Glazebrook, and Weber 2011), in which the team must choose among a small finite set of pre-computed *Free Kick Plans* (FKPs) as their actions, which yield a reward of 1 if they score a goal within a short time after the free kick –e.g., the FKP of Figure 1– or 0 otherwise. The effectiveness of different FKPs heavily depends on the location from which the free kick is taken, so we approach the problem as a contextual multi-armed bandit problem (Dudík et al. 2011) with a metric context (Slivkins 2014).



(a) FKP setup (dashed circle robot passes ball)



(b) FKP execution and scoring (dashed circle robot shoots ball)

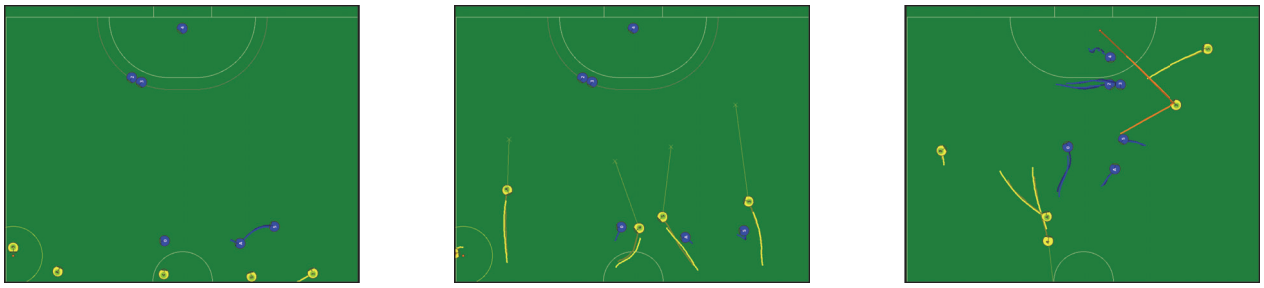
Figure 1: Free Kick Plan (FKP) successfully executed at RoboCup 2015. Yellow circles show our team's robots. We present an algorithm for learning effective FKPs online.

Our proposed approach to learning is for the team to model an estimate of the reward function using Gaussian Process-based regression (Quiñonero Candela and Rasmussen 2005) for each FKP, and then choosing the next action to be the one that maximizes the Upper Confidence Bound (UCB) acquisition function (Srinivas et al. 2009), thus guaranteeing a no-regret learning process.

We evaluate our online learning algorithm using a physics-based SSL soccer simulation. We demonstrate the effectiveness of the algorithm against three realistic defending teams, each with different weaknesses and strengths.

Concretely, this paper presents three contributions: (a) A framework for modeling the problem of online learning of free kicks as a contextual multi-armed bandit problem, (b) An algorithm for addressing this problem, and (c) empirical evidence for the effectiveness of the algorithm.

[1] wiki.robocup.org/wiki/Small_Size_League



(a) Setup: Kicker on the left, others at $\boldsymbol{x}_i^s$ (b) Move to target locations (thin yellow lines) (c) Pass and score (ball path in orange)

Figure 2: Plan fk_5 at RoboCup: Kicker (left yellow) passes as teammates (yellow) charge to the opponent's (blue) goal (top).

Background: Robot Soccer and Free Kicks

The problem of autonomous robot soccer has been investigated by researchers for several years, as it provides a rich domain for a wide range of problems, from computer vision, to walking control and team AI. We focus our research on the SSL, which consists of two teams of 6 wheeled robots that fit within a cylinder with a base with $9cm$ radius and $15cm$ height. These robots play soccer with a golf ball in a field of size $9m \times 6m$, controlled wirelessly by a central computer that gets vision information from the field at 60Hz. The SSL focuses on the problem of team coordination in an adversarial, highly dynamic environment –the robots move at speeds of over $3m/s$ and shoot the ball at up to $8m/s$.

The problem of reinforcement learning has been addressed in several robot soccer applications, such as simulation of keepaway (Stone, Sutton, and Kuhlmann 2005) and layered learning of soccer behaviors (MacAlpine, Depinet, and Stone 2015). However, online learning has not been applied nearly as frequently. Online learning has been applied to adjust weights of different formations depending on their success (Browning et al. 2005), and to improve the models of passing success against an unknown opponent (Mendoza, Veloso, and Simmons 2015); neither of these explicitly addressed the exploration vs. exploitation problem. One of the reasons for the lack of research on online learning for robot soccer is the difficulty of the problem: since soccer games only last a few minutes, and each opponent is only encountered once, robots must learn from very sparse data in a very high-dimensional domain. This is the reason why our work focuses on learning from a small set of actions in the more specific domain of robot soccer free kicks.

In the SSL, free kicks are a method of restarting the game after an infraction, or after the ball has left the field. A free kick is awarded to one of the teams at the closest legal location to the infraction. Until the kicking team restarts play by touching the ball, all robots from the opposing team must maintain a distance of at least $50cm$ from the ball. In this paper, we focus on online learning for offensive free kicks, since they provide semi-controllable initial conditions for our plans –the ball is stationary, and we can choose where to position our robots – and they provide scenarios with a relatively high chance of scoring, especially since we focus on offensive free kicks on the opponent's half of the field.

Free Kick Planning as a Bandit Problem

The full planning space of offensive free kick plays consists of continuous and high-dimensional state and action spaces. The full state space consists of more than 80 physical dimensions, including the position, orientation, and velocities of the 12 robots and the ball, plus the state of the game and the internal state of each team. The action space is also high-dimensional, as robots can move arbitrarily within physical limitations, and they can execute long sequences of passes, dribbling and shooting. We propose to make online learning feasible from sparse observations by greatly reducing the planning space and modeling the problem as a contextual multi-armed bandit problem.

State Space. Offensive free kicks allow our team to control the initial conditions of the world to a large extent: the ball is stationary at position $\boldsymbol{x}^b$, and we can place our robots at arbitrary feasible initial conditions. Furthermore, we assume that the adversary does not change its behavior throughout the game, and thus the opponent's reactions to our plans are relatively repeatable as well. We therefore reduce the planning state $\boldsymbol{s}$ to the two-dimensional initial ball location $\boldsymbol{s} = \boldsymbol{x}^b$ from which the free kick is taken. The effectiveness of different plans highly depends on $\boldsymbol{x}^b$.

Action Space. To reduce the size of the action space, we define a set of *Two-Step Free Kick Plans* (2FKPs) that consist of the following sequence: First, every robot ρ_i, excluding the goalie ρ_g and the free kick taker ρ_k, proceeds to a setup location $\boldsymbol{x}_i^s$; then, the robots proceed to final target locations $\boldsymbol{x}_i^f$, while ρ_k passes to the best potential target robot ρ^*, at the best computed location $\hat{\boldsymbol{x}}_i^f$ within a fixed radius of $\boldsymbol{x}_i^f$. Given a team of N_ρ robots, then, a 2FKP can be expressed as a vector $\boldsymbol{a}$ of length $2(N_\rho - 2)$ of locations $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_i^f$ for each potential receiver. 2FKPs are expressive enough to contain dynamic plans with $\boldsymbol{x}_i^s \neq \boldsymbol{x}_i^f$, yet simple enough to enable a bandit formulation, rather than more general reinforcement learning (Kaelbling, Littman, and Moore 1996) over long sequences of actions.

We further restrict the set of possible free kicks – because of our goal of learning from sparse observations– to a finite set of 2FKPs containing $N_{\boldsymbol{a}}$ elements: $\boldsymbol{A} =$

$\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{N_{\boldsymbol{a}}}\}$. Thus, during each offensive free kick, the team must choose among $N_{\boldsymbol{a}}$ possible actions.

Figure 2 illustrates an example of a 2FKPs being executed during RoboCup 2015. In this fk_5 plan, all the robots except for ρ_g and ρ_k spread around the midfield for the setup (Figure 2a), and then proceed to charge forward to locations around the opponent's goal (Figure 2b) to receive a pass and shoot (Figure 2c).

Reward Function. We seek to maximize the number of goals scored during offensive free kicks, and thus we specify the rewrad function r as $r = 1$ if our team scores within time t_{FK} of the kick, or $r = 0$ otherwise. Time t_{FK} is a threshold indicating an approximate time after which scoring is no longer attributed to the chosen 2FKP; in our work, $t_{FK} = 10s$.

Online Learning over a set of 2FKPs

We enable online learning by (1) modeling an approximation to the expected reward function $\bar{r}(\boldsymbol{s}, \boldsymbol{a})$ as a function of the state of the world $\boldsymbol{s}$ and the chosen action $\boldsymbol{a}$, and (2) appropriately choosing actions that intelligently trade-off exploitation of known good actions, with exploration of actions with uncertain results.

Approximating $\bar{r}$ using Gaussian Processes We approximate $\bar{r}$ using Gaussian Processes (GPs)-based regression (Quiñonero Candela and Rasmussen 2005). We do this through a bank of GPs $\{\text{GP}^{\boldsymbol{a}_1}, \text{GP}^{\boldsymbol{a}_2}, \ldots, \text{GP}^{\boldsymbol{a}_{N_{\boldsymbol{a}}}}\}$, one for each available action. We start by approximating $\bar{r}$ by a prior function $\hat{r}_0^{\boldsymbol{a}}(\boldsymbol{s})$; then, we use subsequent observations $(\boldsymbol{s}^i, \boldsymbol{a}^i, r^i)$ of free kick states $\boldsymbol{s}^i$, chosen action $\boldsymbol{a}^i$, and resulting reward r^i to update this estimate online. Thus, the approximating function $\hat{r}(\boldsymbol{s}, \boldsymbol{a})$ is given by:

$$\hat{r}(\boldsymbol{s}, \boldsymbol{a}) = \text{GP}^{\boldsymbol{a}}\left(\boldsymbol{s} | \hat{r}_0^{\boldsymbol{a}}, (\boldsymbol{s}^1, r^1), \ldots (\boldsymbol{s}^n, r^n)\right),$$

where each pair $(\boldsymbol{s}^i, r^i)$ is an observation in which $\boldsymbol{a}^i = \boldsymbol{a}$. Our algorithms can query this bank of GPs to find the expected value $\mu(\boldsymbol{s}, \boldsymbol{a})$ and the variance $\sigma^2(\boldsymbol{s}, \boldsymbol{a})$ of the estimate expected reward function $\hat{r}(\boldsymbol{s}, \boldsymbol{a})$. Figure 3 shows the expected value function mu estimated from success and failure data of a particular 2FKP.

(a) Reward: 0 (black), 1 (white) (b) Reward function estimate

Figure 3: Example expected reward estimate over valid free kick locations. We only show samples from one half of the field length-wise, as our implementation assumes symmetry.

Upper Confidence Bound Online Learning Given estimates for the mean and variance of the modeled reward function, we use the Upper Confidence Bound (UCB) algorithm to choose the next action $\boldsymbol{a}$, given state $\boldsymbol{s}$:

$$\boldsymbol{a}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}' \in \boldsymbol{A}} \left[\mu(\boldsymbol{s}, \boldsymbol{a}') + \beta\sigma(\boldsymbol{s}, \boldsymbol{a}')\right], \quad (1)$$

where β is a parameter that controls the level of exploration vs. exploitation in the algorithm. The UCB algorithm has been shown to be a no-regret algorithm (Srinivas et al. 2009), guaranteeing that difference between the reward of our chosen action and the reward of the optimal action grows sublinearly as the team learns which action to take.

Algorithm 1 illustrates the process of online learning of FKPs during a game. If the team must select a free kick plan to execute, it uses Equation 1 to choose the next action. At the end of the play, the team sees a reward of either 0 or 1, depending on whether it scored a goal, and adds the observation to the right GP.

Algorithm 1 Free kick plan online learning procedure.

```
procedure LearnFreeKick(game_state)
    if game_state = select_free_kick then
        s_i ← x^b
        a_i = arg max_{a∈A} [μ(s_i, a) + βσ(s_i, a)]
    end if
    if game_state = play_end then
        r_i ← 1 if goal, 0 otherwise
        Add (s_i, r_i) to GP^{a_i}
    end if
end procedure
```

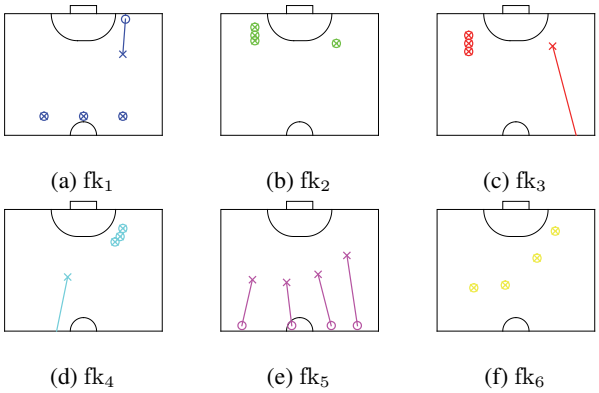
Experimental Results

The goal of this work is to achieve advantageous FKP adaptation online during RoboCup games. In fact, during RoboCup 2015, our team did perform such adaptation during real games, and it won the SSL tournament. However, it is very difficult to accurately evaluate the amount of credit that the online learning of free kicks deserves in that victory due to (i) lack of ground truth, (ii) small number of games, and (iii) the large proportion of the games that does not involve free kicks, such as offense coordination during regular gameplay (Mendoza et al. 2016a). Thus, we instead present a controlled experimental evaluation of our algorithm.

We evaluate the proposed FKP modeling and action selection algorithm on a PhysX-based simulation of a robot soccer game of the SSL. We equipped the team with 6 different 2FKPs, illustrated in Figure 4. Here, we describe how we obtained the true expected reward function of each 2FKP, and the results of evaluating our algorithm against three different defending teams.

Defense Teams

We evaluated our online learning algorithm against three different defense teams. For each defense, the closest robot to the goal is the goalie, who intercepts incoming shots, and the closest robot to the ball attempts to gain control of it by



(a) fk_1 (b) fk_2 (c) fk_3

(d) fk_4 (e) fk_5 (f) fk_6

Figure 4: Available free kick plans. Circles mark initial receiver locations $\boldsymbol{x}_i^s$, while Xs mark their target locations $\boldsymbol{x}_i^f$. Plots assume that $\boldsymbol{x}^b$ is in the left half of the field; otherwise, the plans are mirrored about the x-axis.

driving to intercept it optimally. One or two of the remaining robots, depending on how many are needed, block open angles from $\boldsymbol{x}^b$ to the goal.

The remaining robots are assigned to block threats on the goal from the most threatening opponents. This evaluation of threats is different for each defense. The *Time Defense* ranks opponent threats based on how long it would take each robot located at location $\boldsymbol{x}^i$ to receive a pass from $\boldsymbol{x}^b$ to $\boldsymbol{x}^i$, assuming a ball speed of $5m/s$, and then shoot on the goal, assuming the maximum legal shot speed of $8m/s$, prioritizing robots with shorter times. The *Angle Defense* ranks opponent threats based on the size of the open angle they have on the goal from their location $\boldsymbol{x}^i$, prioritizing robots with wider shooting angles. The *Combined Defense* ranks opponent threats based on a combination of the measures of Time and Combined, using the open angle measure of Angle only if the robot's open angle is smaller than a threshold $\phi_{\max}$. If multiple robots have an open angle wider than $\phi_{\max}$, they are compared based on the time measure of Time. This realistic defense threat evaluation is the actual evaluation used by the CMDragons SSL team (Biswas et al. 2014), who are the current champions of the SSL (Mendoza et al. 2016b).

Online Learning Evaluation

To evaluate our algorithm against the 3 different teams above, we first obtain an accurate estimate of the expected reward function $\bar{r}^{\boldsymbol{a}}$, for each action $\boldsymbol{a}$, and for each of the 3 teams. Then, using online learning, we evaluate the evolution of the expected regret in time.

True Expected Reward. For each 2FKP, we ran extensive simulation free kicks from a fine grid of locations $\boldsymbol{x}^b$, and used GPs to model the expected reward. This function approximation becomes increasingly accurate as we run more free kicks; we obtained the true reward function with ~ 1000 free kicks for each 2FKP. Figure 5 illustrate the resulting expected reward function of the optimal action $\boldsymbol{a}^*$ for each free kick location $\boldsymbol{x}^b$ on the field.

Online Learning Performance. We evaluated our algorithm against each defense by conducting sequences of free kicks from the forward-left quadrant as during training, but using a random sequence instead of a grid sequence. Furthermore, the team selected their action according to our online learning algorithm. For each chosen action, we measured the expected regret R as:

$$R(\boldsymbol{a}^i|\boldsymbol{s}^i) = \max_{\boldsymbol{a}' \in \boldsymbol{A}} \left[\bar{r}^{\boldsymbol{a}'}(\boldsymbol{s}^i) - \bar{r}^{\boldsymbol{a}^i}(\boldsymbol{s}^i) \right] \quad (2)$$

Figure 6 illustrates the average evolution of regret for each defense team, along with the expected reward of the optimal action and the expected reward of the chosen action. The optimal reward, as we average over more learning episodes, converges toward a horizontal line, different for each of the three defenses. The regret decreases significantly even within the first 10-20 free kicks, and for the Time defense it nearly reaches 0 within that time; this indicates that our algorithm could significantly improve the performance of our team during RoboCup games.

Conclusion

This paper presents an algorithm for modeling and online learning of Free Kick Plans (FKPs) in games of autonomous robot soccer. To achieve our goal of enabling robots to adapt to different opponents within the time scale of a game of soccer, we greatly reduce the size of (a) the state space, to the two-dimensional position of the free kick and (b) the action space, by creating a small finite set of Two-Step Free Kick Plans. During execution, our algorithm models the expected reward function of the different actions, given the state, using Gaussian Processes. To adapt online, the robots use the UCB algorithm to select their actions, effectively trading off exploitation with exploration. Empirical results demonstrate that our team was able to adapt, using a small number of data points, to various opponents, including a realistic defense modeling that of a successful SSL team.

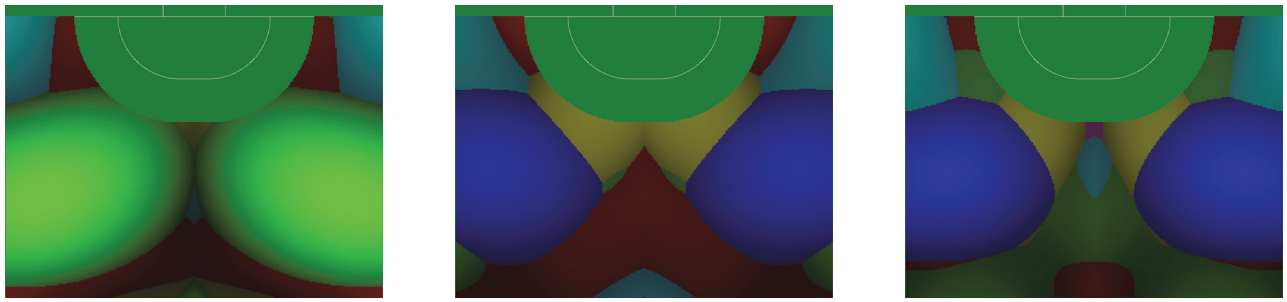
Acknowledgments

This material is based on work partially supported by NSF Grant IIS-1012733, DARPA Grant FA87501220291, and MURI subcontract 138803 of Award N00014-09-1-1031. The presentation reflects only the views of the authors. This material is based upon research supported by (while Dr. Simmons was serving at) the National Science Foundation.

References

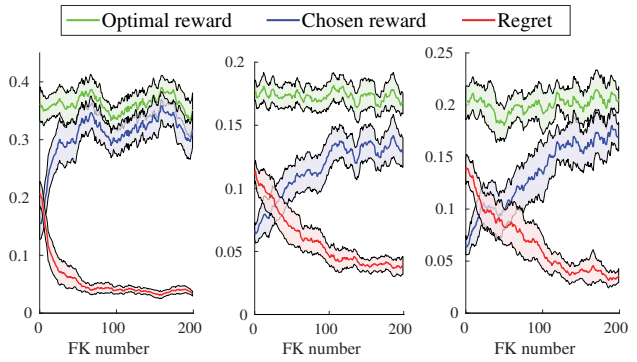
Biswas, J.; Mendoza, J. P.; Zhu, D.; Choi, B.; Klee, S.; and Veloso, M. 2014. Opponent-driven planning and execution for pass, attack, and defense in a multi-robot soccer team. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Browning, B.; Bruce, J.; Bowling, M.; and Veloso, M. 2005. STP: Skills, tactics, and plays for multi-robot control in adversarial environments. *Proceedings of the IME, Part I: Journal of Systems and Control Engineering* 219(1):33–52.



(a) Time defense (b) Angle defense (c) Combined defense

Figure 5: Optimal action map. Different colors represent different FKPs, consistent with the colors of Figure 4. Color intensity, between 0 and 1, is the nonlinear function $\sqrt{\bar{r}}$ for ease of visualization.



(a) Time defense (b) Angle defense (c) Combined defense

Figure 6: Results of online learning against three defense teams. Average regret decreases quickly for each of the defenses. Shaded areas indicate a 95% confidence interval for the value of each function.

Dudík, M.; Hsu, D.; Kale, S.; Karampatziakis, N.; Langford, J.; Reyzin, L.; and Zhang, T. 2011. Efficient optimal learning for contextual bandits. *CoRR* abs/1106.2369.

Gittins, J.; Glazebrook, K.; and Weber, R. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* 4:237–285.

MacAlpine, P.; Depinet, M.; and Stone, P. 2015. UT Austin Villa 2014: RoboCup 3D simulation league champion via overlapping layered learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, volume 4, 2842–48.

Mendoza, J. P.; Biswas, J.; Cooksey, P.; Wang, R.; Klee, S.; Zhu, D.; and Veloso, M. 2016a. Selectively reactive coordination for a team of robot soccer champions. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI)*.

Mendoza, J. P.; Biswas, J.; Zhu, D.; Wang, R.; Cooksey, P.; Klee, S.; and Veloso, M. 2016b. Cmdragons 2015: Coordinated offense and defense of the ssl champions. In *RoboCup 2015: Robot World Cup XIX*.

Mendoza, J. P.; Veloso, M.; and Simmons, R. 2015. Detecting and correcting model anomalies in subspaces of robot planning domains. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (to appear)*.

Quiñonero Candela, J., and Rasmussen, C. E. 2005. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.* 6:1939–1959.

Slivkins, A. 2014. Contextual bandits with similarity information. *The Journal of Machine Learning Research* 15(1):2533–2568.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. W. 2009. Gaussian process bandits without regret: An experimental design approach. *CoRR* abs/0912.3995.

Stone, P.; Sutton, R. S.; and Kuhlmann, G. 2005. Reinforcement learning for RoboCup-soccer keepaway. *Adaptive Behavior* 13(3):165–188.