

Beyond action valuation: A deep reinforcement learning framework for optimizing player decisions in soccer

Pegah Rahimian[†] – Jan Van Haaren[‡] – Togzhan Abzhanova – Laszlo Toka[†]

[†]Budapest University of Technology and Economics, Hungary – {pegah.rahimian,toka}@tmit.bme.hu

[‡]Club Brugge-KU Leuven, Belgium – jan.vanhaaren@kuleuven.be

1. Introduction

Soccer players need to make many decisions throughout a match in order to maximize their team's chances of winning. Unfortunately, these decisions are challenging to measure and evaluate due to the low-scoring, complex and highly dynamic nature of soccer. Most decisions have little immediate impact but may positively contribute to the team winning in the long run. For instance, a simple short pass in midfield may open up valuable space elsewhere on the pitch for a teammate. Like Johan Cruyff once said: *Sometimes something's got to happen before something is going to happen*. In the past few years, several different methods have been proposed to value the on-the-ball actions that players perform as a result of their decisions [1,2,3,4,5]. However, very little work has been done on evaluating the alternative actions that a player or team could have performed in each situation in order to increase the team's chances of winning [6,7,8].

To overcome the limitations of existing action valuation metrics, we propose a framework that evaluates each decision that a player makes with respect to passing or shooting the ball in each game state and suggests the optimal decision that the player could have made in that game state. Concretely, we represent team behavior using two pitch surfaces: a *selection probability* surface that reflects the likelihood of passing the ball to each pitch location from the current game state, and a *success probability* surface that reflects the likelihood of a pass to each pitch location being successful. Using tracking data that provides the locations of all 22 players and the ball, we capture a team's actual behavior as a Markov Decision Process (MDP) and use Reinforcement Learning (RL) to obtain a team's optimal behavior. We identify situations where players and teams could have made better decisions by comparing their actual behavior to their team-specific optimal behavior.

Figure 1 demonstrates a potential application of our proposed framework. The visualization shows a possession from Cercle Brugge in a game against Union Saint-Gilloise in the 2021-22 season of the Belgian Pro League [20]. The chart shows how the Expected Possession Outcome (EPO) evolves throughout the possession when following the actual behavior and the optimal behavior that our framework derived. The EPO ranges from -1 to +1, where +1 corresponds to scoring a goal at the end of a possession and -1 corresponds to conceding a goal at the end of a possession.

Our optimization framework enables soccer clubs to analyze their pass and shot tendencies on different pitch zones and different phases of ball possession. Moreover, they can compare their team-specific action selection tendencies with the optimal ones that lead them to beat the opponent in a game. It also allows players and coaches to discover the optimal ball destination and pass direction for any game situation by maximizing the expected outcome of all possessions throughout the game. This paper is organized as follows. Section 2 describes our dataset. Sections 3 and 4 explain our approach to deriving the actual and optimal team behavior. Section 5 discusses the evaluation of the

optimal team behavior. Section 6 presents several use cases demonstrating the practical applicability of our framework. Section 7 discusses the related work. Section 8 concludes our work.

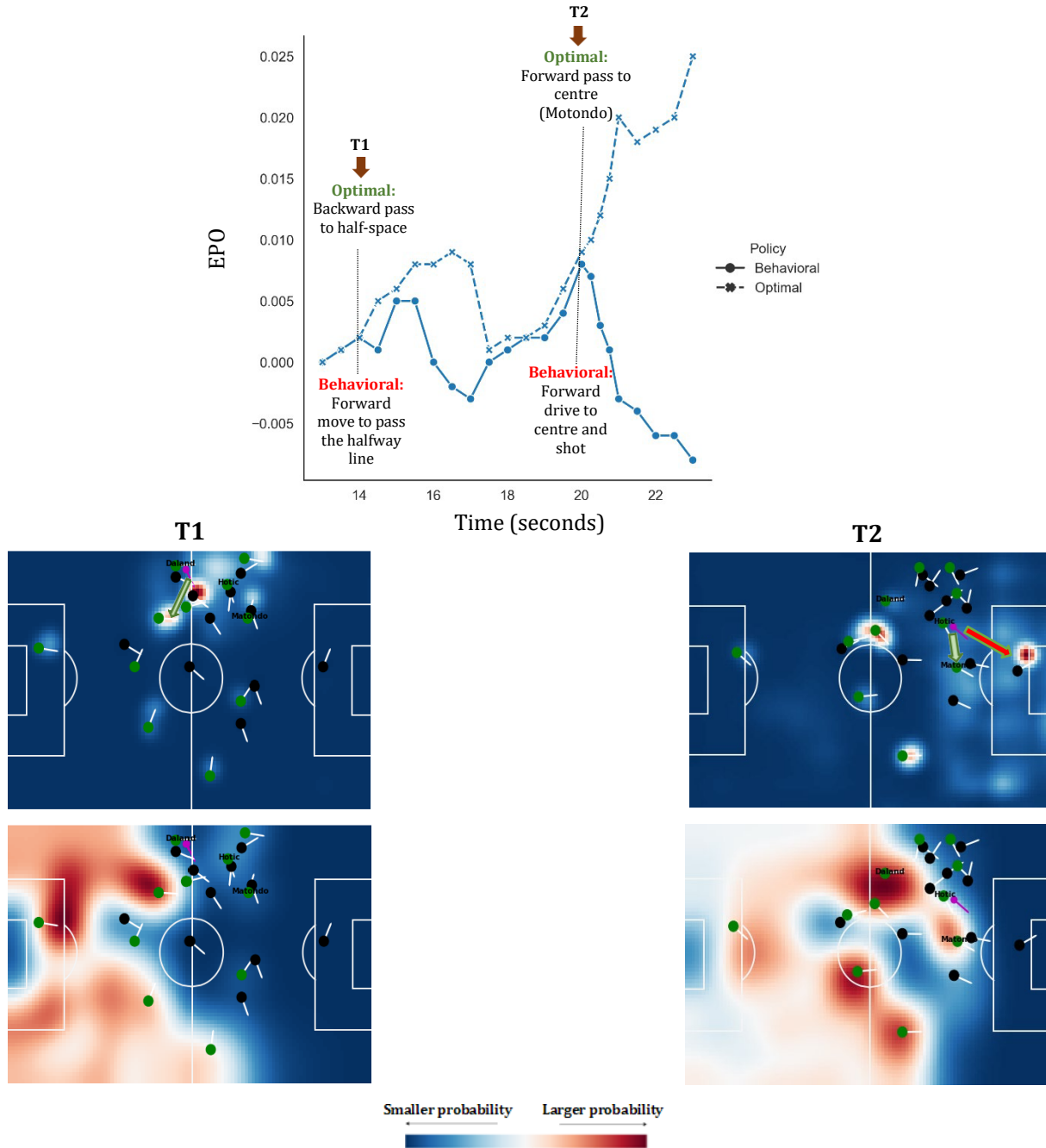


Figure 1. A Cercle Brugge possession evaluation in a game against Union Saint-Gilloise in the 2021-22 season of the Belgian Pro League. The top pitch visualizations show the selection probability surfaces according to the actual behavior. The green circles indicate the Cercle Brugge players, the green and red arrows indicate the optimal and actual ball movement, respectively. The pitch visualizations at the bottom show the reward for moving the ball to each pitch location. (Link of the video¹)

¹ bit.ly/3HC2j5X

2. Dataset description

The dataset consists of high-resolution spatiotemporal tracking and event data covering all 330 games of the 2020-21 season, and 100 games of 2021-22 season of Belgian Pro League collected by Stats Perform until the submission of this paper, i.e., October 2021. The tracking data include the (x,y) coordinates of all 22 players and the ball on the pitch at 25 observations per second. The event data includes on-ball action types such as passes, shots, dribbles, etc. annotated with additional features such as contestants, period ID, ball possessor player ID, start and end locations of the ball. We then merged tracking with event data. Each record of our merged dataset includes all players and the ball coordinates with their corresponding features for each snap-shot, i.e., every 0.04 of a second. Figure 2 illustrates a three-second of our dataset during the Cercle Brugge-Union SG match in October 2021.

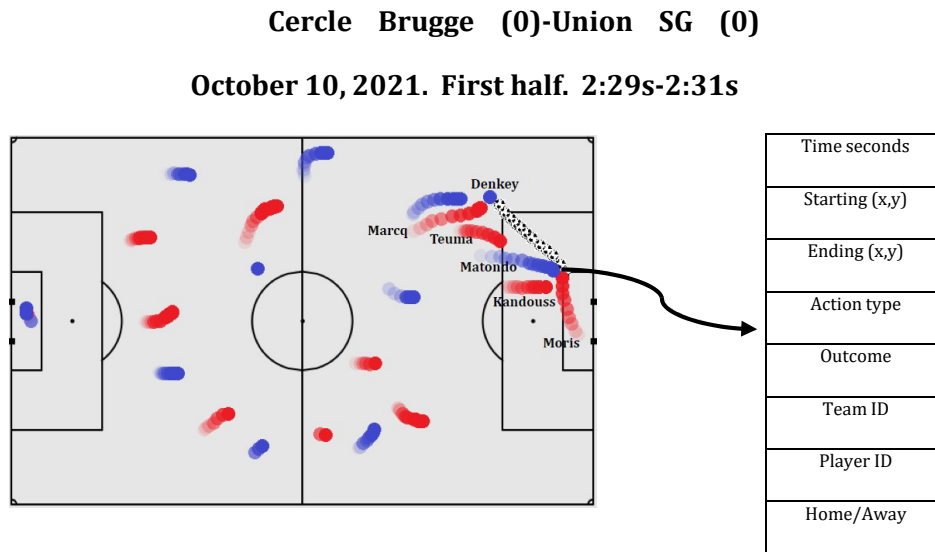


Figure 2. Three-second of the Cercle Brugge (blue dots)-Union SG (red dots) match on 10 October 2021. Denkey possesses the ball at 2:29s, deciding about the ball destination. He sends the ball towards the goal area, where Matondo attempts an unsuccessful shot. Meanwhile, Kandouss scrambles to defend him, and Moris saves the ball.

3. How to model the actual behavior of a team?

In this section, we elaborate on the technical characteristics of deriving a behavioral policy on both the league-level and the team-level. The policy in this work is defined as the team propensities of selecting the ball destination, i.e., a location on the pitch, given the positions of the 22 players and the ball. For instance, considering the first snapshot of Figure 2, our trained policy estimates the probability of ball carrier Denkey to pass the ball to each location on the pitch. In order to represent these probabilities for any situation of the game and use them for further decision making (e.g., where should Denkey send the ball?), we need to estimate two probability surfaces: The first probability surface is the **selection surface**, which shows the probability of the ball being passed to each pitch location from a given game state. The second probability surface is the **success surface**, which predicts the probability of the action being successful (i.e., possession is maintained) for each location on the field if the ball is sent over to that location. Figure 5 visualizes the surfaces for a particular

game situation. These surfaces are obtained by carefully training a policy network that receives a particular game situation as input, and produces the probability surfaces as outputs. In order to include the effects of the game context in the model, we build the policy network on top of the event- and tracking data, and we use deep learning techniques to tackle the complexity of spatiotemporal tactics. Now we describe the game state representation, and the architecture of the policy network to obtain the required surfaces.

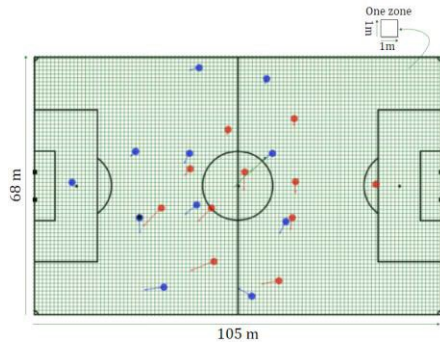
3.1. State representation and input channels

We represent the game state for each snapshot of our dataset described in Section 2, including the specific locations (x,y) of all players and the ball, their corresponding velocities, and respective outcomes of particular events (e.g., success or turnover for pass). To represent this information, we construct eleven input channels in a format that suits the policy network. Each channel is a matrix of size (105×68) , representing the length and width of the game field respectively (see Figure 3). The input data channels contain different types of low-level information for each square meter of the field (one disjoint zone in Figure 3) to obtain a contextual representation of the game at a given time step. We constructed the following input channels to represent each game state:

1. **$InChI_t$** : locations of the attacking team's players. The value of every player's location is set to 1.
2. **$InChI_o$** : locations of the defending team's players. The value of every player's location is set to 1.
3. **$InChV_{xt}$** : x components of the velocities of the attacking team's players.
4. **$InChV_{yt}$** : y components of the velocities of the attacking team's players.
5. **$InChV_{xo}$** : x components of the velocities of the defending team's players.
6. **$InChV_{yo}$** : y components of the velocities of the defending team's players.
7. **$InCh_{db}$** : Euclidean distance of the ball to each location on the field.
8. **$InCh_{dg}$** : Euclidean distance of the defending team goal to each location on the field.
9. **$InCh_{abg}$** : angle between the ball and the goal for each location on the field.
10. **$InCh_{cosabg}$** : cosine of the angle between the ball and the goal for each location on the field.
11. **$InCh_{sinabg}$** : sine of the angle between the ball and the goal for each location on the field.

Fine-grained partitioning of the pitch:

Contains $105 \times 68 = 7140$ disjoint zones



State representation:

(Input shape for a given state to the policy networks)

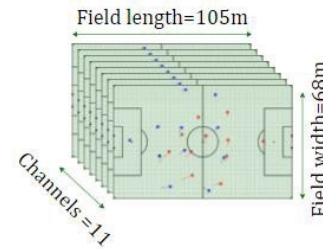


Figure 3. State representation. Blue, red, and black circles represent the home team players, away team players, and the ball, respectively. Each state represents a time step in the dataset, containing 11 matrices with size 105×68 , each for an input channel covering all 1×1 zones.

3.2. The policy network architecture

To infer our probability surfaces and to formulate the policies, we apply deep learning techniques. We use a policy network, which is a neural network that takes a huge number of game states as input and produces the probability surfaces as outputs.

From a technical perspective, any probability surface contains $105 \times 68 = 7140$ probabilities, one for each disjoint zone on the pitch given a specific game state. In the selection surface the sum of probabilities over all zones adds up to 1. But the probabilities for each zone of the success surface represent the likelihood of the actions being successful (i.e., possession is saved for the team) if the ball is sent over to that zone. Note that the estimated selection surface shows the behavioral policy of the teams in terms of selecting the ball destination location, given the current locations of all players and the ball. Figure 4 illustrates the architecture of the policy network producing the above-mentioned probability surfaces. The input frame sequences are the input channel matrices described in the previous section. Outputs 1 and 2 show two probability surfaces, namely *selection surfaces* and *success surfaces*. The output of the policy network depends on the setup of the last layer of the encoder: setting the activation function to softmax and sigmoid would yield the selection and success surfaces, respectively. Elaborating on the choice of layers, the up and down-sampling, and the loss functions in the different layers are beyond the scope of this paper. This paper focuses on the practical application of the network rather than the technical characteristics and setup.

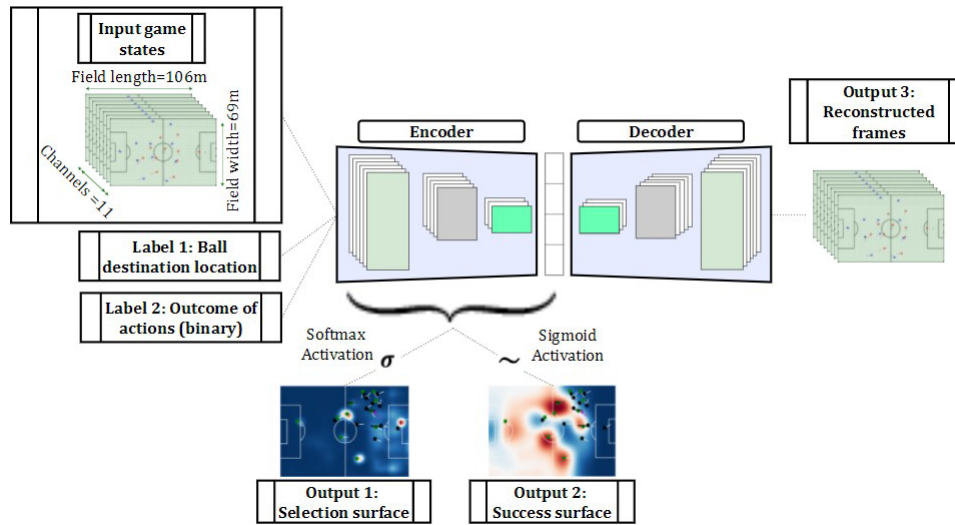


Figure 4. Policy network architecture with the respective inputs and outputs. Output 1 is the selection probability surface that is considered as the policy in this work. Output 2 is the success surface representing the probability of action being successful if the ball is sent over to each location on the field. Output 3 is the reconstructed sequence of actions (a new possession) to simulate the learned policy of the network.

Table 1 shows the results of the selection and success probability estimation of the policy network for all passes in the test set. We used a split of 80%-10%-10% for the train-validation-test sets respecting the chronological order of the games. To evaluate the performance of our trained policy network, we compare the results with the following three baseline models. The first baseline is the

Naïve model only for the success estimation proposed by [9], which assigns the average pass completion of 85% to all passes. Unfortunately, we could not come up with a similar Naïve model for the selection estimation. The second and third baselines are traditional classifiers (i.e., logistic regression and XGBoost) to predict the selection and success probabilities of each pass. Although using machine learning techniques (e.g., logistic regression) is quite efficient when evaluating the individual passes in the game, its usage can be cumbersome when it comes to estimating the full probability surfaces, which requires setting 105×68 classes for the selection surface, and 105×68 different binary classifications for each success surface. As Table 1 shows, the baseline models cannot capture the complex intricacies of the games and yield higher log-loss in comparison to our proposed model. Although applying hand-crafted features, such as team tactics and formations as suggested in [9], can considerably decrease the log-loss, its long inference time makes the baseline model impractical for our analysis. The hyper-parameters of all models are optimally selected using the validation dataset.

Table 1. Training result of the models on all passes in the test set.

Surface type	Model	Log-loss	Inference time	Parameters
Selection	Logistic regression	0.68	151 seconds	12
	XGBoost	0.65	127 seconds	-
	Proposed network	0.18	7.78 seconds	431,354
Success	Naïve	0.45	-	-
	Logistic regression	0.57	372 seconds	12
	XGBoost	0.58	351 seconds	-
	Proposed network	0.22	2.27 seconds	331,276

Figure 5 illustrates the surface outputs of our trained policy network on a specific game state during a match of Cercle Brugge against Union SG. The estimated probability surfaces illustrate the performance of our policy network on capturing the influence of teammates, opponents, and their velocities on the ball holder (Hotic) on deciding about the ball destination.

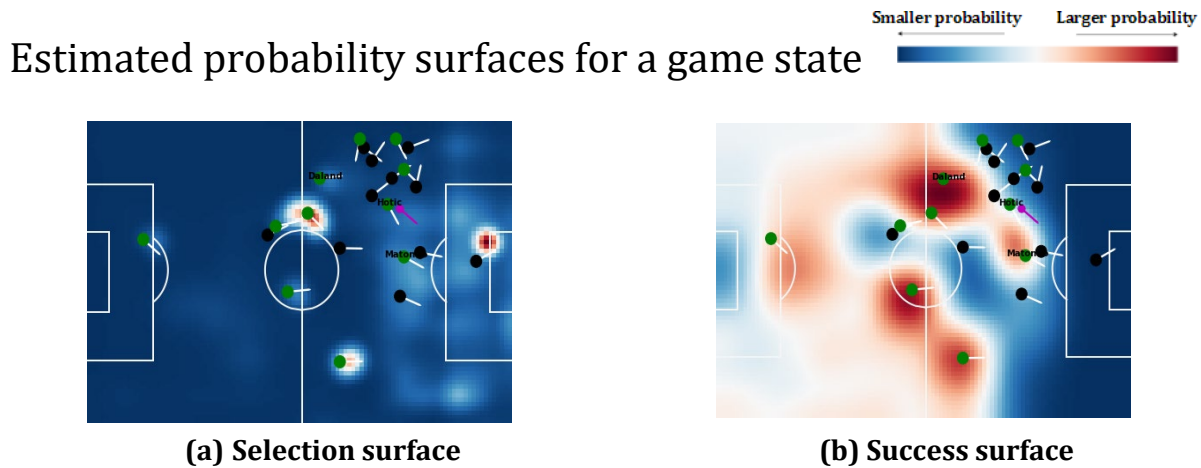


Figure 5. Output surfaces for a game state in a match of Cercle Brugge against Union SG. The surface color represents the probabilities from cool to warm, where cool (blue) represents lower probability, and warm (red) represents higher probability. Green and black circles represent the players' locations on the attacking and defending team, and the arrows represent their velocities. The purple circle represents the ball's location.

4. How to obtain the optimal behavior of a team?

So far, we have prepared the selection and success probability surfaces for each game state. But the estimated surfaces are not comprehensive enough to assist the players and coaches in optimal decision making: the selection surfaces are estimated according to the general policy of the historical games, and there is no evidence of optimality of the decisions and policies made by teams and players in prior matches. Moreover, the estimated success surfaces only indicate the short-term rewards of not losing the possession that the player can gain if he/she moves the ball to that location. The current analytical methods [6,10] propose estimating value surfaces by training the neural network to predict the probability of goal scoring within the next 10 actions or at the end of the actual possession. In this section, we elaborate on our proposed optimization algorithm that can directly estimate the optimal full probability surfaces, covering all ball destination locations on the field, rather than learning values for each of the discrete actions that occurred in prior games.

4.1. Markovian possession environment

A Markov Decision Process (MDP) is a framework used in systems with an urge for sequential decision making. When it comes to soccer ball possessions, the sequential nature of actions within a possession allows us to model possessions with this popular technique for optimization in RL tasks. The MDP models the probability that the ball carrier selects the destination location of the ball on the field, given the current positions of all 22 players and the ball. Our approach of modeling soccer with an MDP requires a number of well-defined elements: a tuple of (S, A, R, π) , where S represents the set of states, A represents the set of actions, R represents the reward function formulating the reward the agent receives for any given state/action pair, and π represents the policy that is interpreted as the probability that the agent takes any given action based on the current state of the environment. Now we can define each of these components in our soccer possession environment:

Episode (τ): Each episode begins from the first event in the ball possession by the team in possession (i.e., transition phase) and culminates in either a goal being scored, or the possession being lost and transferred to the opponent. Since there is no unified definition of a possession in soccer, we define the possession as: a sequence of actions from the beginning of a deliberate on-ball action by a team, until they score a goal or lose the ball to the opposing team for more than two consecutive actions. Thus, when opposing players touch the ball fewer than three consecutive times, we do not consider that interruption as a loss of possession.

State (s): The state consists of the players' and ball's locations and the rest of the previously described input channels (e.g., players' velocities, distances, and angles to goal) for any game situation. There are two absorbing states: goal scoring, and loss of possession. Rebounds after unsuccessful shots, balls going out of play, and fouls are not considered as a loss of possession if the ball is recovered by the team within the next three actions.

Action space (a): We use two types of action space in this paper. First, we consider a continuous action space, where the action is defined as the specific location on the field, selected by the ball carrier, to move the ball to. Second, we discretize the action space into backward pass, forward pass, sideway pass, and shot, for interpretability and explainability reasons.

Policy (π): We define the policy as the probability with which the ball carrier selects any specific location on the field as the destination location of the ball, given the current state.

Reward signal ($R(s, a)$): The main reward in soccer comes from winning the game. However, such a reward signal is far too sparse for an agent learning to act optimally. Thus, we need to handcraft

the reward function to encourage relative behaviors in distinct phases/states of the game. Due to the different objectives of the distinct phases of a soccer game (i.e., non-stationary policy), our reward engineering approach applies various reward functions in different phases with the aim of incorporating the tactical characteristics of the game in their design. We split each possession into several phases of play according to the Opta possession framework² (see Figure 6) and tag the events according to their locations on the field, and what happened before and after them.

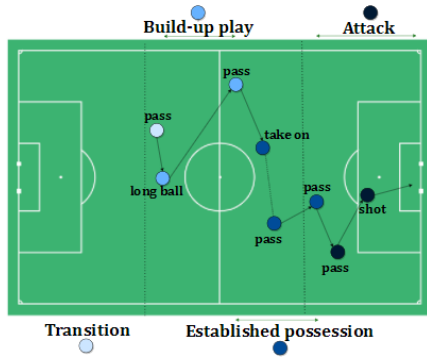


Figure 6. Different phases of possession:

- **Transition:** Period between regaining possession and moving it away from pressure.
- **Build-up play:** Events originating from a team's own half moving the ball into midfield, either centrally or out wide. This phase ends when the ball crosses the halfway line.
- **Established possession:** Having possession under control in the middle third by players taking a couple of touches or playing a high number of quick passes between players.
- **Attack:** Having controlled possession in the attacking third.

We assign a positive reward to successful actions (i.e., actions that kept possession) and a negative reward to unsuccessful actions (i.e., actions that led to a loss of possession). While an arbitrary choice, the negative reward is needed to discourage the agent from repeating the same action in the same situation in the future. For the unsuccessful actions, we use the negative of the expected-goals value for the opponent's shot, where the expected-goals value corresponds to the likelihood of the shot resulting in a goal [11]. For the successful actions, we use a tailored positive reward function for each of the four different possession phases:

1. **Transition phase:** From start of the possession (i.e., transiting from defense to attack) until the player completes the first pass (successfully) or loses the ball.
 - Objective: Move the ball away from contact and change the horizontal channel³ (i.e., areas created between the defense, midfield, and attack).
 - Reward function: First we cluster the opponents into three pressure clusters using the K-Means algorithm according to their locations and velocities. Then we assign higher rewards to the actions that move the ball further away from the cluster centroids, and to a location with higher success probability ($P_{success}$).

$$r_1(s, a) = \begin{cases} P_{success}(x, y) \times \sum_{i=1}^3 d_i((x, y), centroid_i) & \text{if possession is retained;} \\ -xG_{opponent} & \text{else,} \end{cases}$$

where (x, y) is the destination location of the ball, $\sum_{i=1}^3 d_i((x, y), centroid_i)$ is the sum of the normalized Euclidean distances between the destination location and the cluster centroids, and $xG_{opponent}$ is the expected goals value for the opponent team.

² <https://www.statsperform.com/resource/phases-of-play-an-introduction/>

³ [https://en.wikipedia.org/wiki/Channel_\(association_football\)](https://en.wikipedia.org/wiki/Channel_(association_football))

2. **Build-up play phase:** Start from playing in their own half (to include goalkeeper and defenders) until the ball reaches the opposition half.
 - Objective: Looking for opportunities to break through the midfield line of the opponent team.
 - Reward function: Depends on the score difference at the given moment of the game. If the attacking team's score is less than or equal to the defender's, the objective is to move to the attack phase as soon as possible and create a chance for scoring. Thus, moving the ball to the location on the pitch with the most success probability and closer to the halfway line should be better rewarded. In case the attackers' score is higher than the defenders', they tend to just keep the possession for as long as they can, while scoring is not the highest priority. Thus, we only consider the probability of success as the reward for those actions.

$$r_2(s, a) = \begin{cases} \frac{P_{success}(x, y)}{d((x, y), \text{halfway line})} & \text{if possession retained and } S_{attackers} \leq S_{defenders}; \\ P_{success}(x, y) & \text{if possession retained and } S_{attackers} > S_{defenders}; \\ -xG_{opponent} & \text{else,} \end{cases}$$

where (x, y) is the destination location of the ball, $S_{attackers}$ and $S_{defenders}$ are the scores of the current game state for the attacking and defending teams, and $d((x, y), \text{halfway line})$ is the normalized horizontal distance of the ball destination location from the halfway line.

3. **Established possession phase:** From the first pass in the opposition's half until the final third of the pitch with over two consecutive actions.
 - Objective: Retain possession (i.e., avoid possession loss).
 - Reward function: We assign larger reward to the actions moving the ball to the location on the pitch with the most success probability.

$$r_3(s, a) = \begin{cases} P_{success}(x, y) & \text{if possession is retained;} \\ -xG_{opponent} & \text{else,} \end{cases}$$

4. **Attacking play phase:** Having controlled possession in the attacking third.
 - Objective: Create chance and goal scoring
 - Reward function: We assign a larger reward to the actions moving the ball to the location on the pitch with a higher success probability and a larger expected goals value.

$$r_4(s, a) = \begin{cases} P_{success}(x, y) \times xG & \text{if possession is retained;} \\ -xG_{opponent} & \text{else,} \end{cases}$$

where xG is the expected goals value of the attacking team that evaluates the shot quality, given the current shooting location and all players' locations on the field [11]. In order to estimate xG for each location on the pitch, we use the capability of the policy network by passing through the input channels of all shots in our dataset, and their labels, i.e., whether they ended in a goal or not. By customizing the last layer of the policy network, it can calculate the xG given any state of the game.

4.2. Expected possession outcome

So far, we have assigned a reward to each action in the dataset according to its phase of occurrence during a possession. However, the action with the highest reward is not necessarily the optimal action that the player could perform, as the assigned rewards estimate only the short-term success and do not consider what will happen at the end of the possession (i.e., scoring a goal or losing possession). In order to address this issue, we introduce the notion of Expected Possession Outcome (EPO) for which we took inspiration from discounted rewards in RL algorithms. EPO is a real-valued number in the range $(-1,1)$ after normalization. We interpret the value as the likelihood of the respective possession ending in a goal for the attacking team (1) or a goal for the opposing team (-1). Assuming a dataset containing N episodes, $D = \{\tau_1, \tau_2, \dots, \tau_N\}$, and an H -step episode (a possession consisting of H actions), $\tau_i = \{s_i^t, a_i^t, r_i^t\}_{t=1}^H$, in which the actions follow policy π estimated by the policy network, the EPO can be formulated as in (1):

$$EPO = \sum_{t=1}^H \mathbb{E}_{s_t, a_t \sim \pi} [\gamma^t r(s_t, a_t)] \quad (1)$$

where $r(s_t, a_t)$ is the assigned reward for the continuous action a_t (i.e., the destination location of the ball on the field) given state s_t according to its phase of occurrence described in the previous section, and γ is the discount factor that we set to 0.99 in this work after carefully tuning with the discount trap method proposed in [12]. The EPO can be interpreted as follows: the strength of encouraging a sample action is the weighted sum of rewards afterwards through a possession. If the possession continues until the attacking phase, the reward of the last action is associated with xG . Thus, EPO shows the likelihood of a possession ending in a goal and can be regarded as the objective function of our optimization framework to be maximized.

4.3. Policy Gradient: what to do on the pitch to maximize the number of goals?

In order to maximize the likelihood of possessions ending in a goal, we need to determine the actions that maximize (1) for all matches in our dataset. According to the fact that carrying out real-world experiments to find the optimal policy would be next to impossible, RL helps us to seek for optimal solutions. To do so, we take advantage of the Policy Gradient (PG) algorithm [13,14,15], that is a popular technique in RL to get the optimal actions in a continuous space. As we aim to estimate the optimal selection surface, the PG algorithm is a perfect choice for optimization: our soccer analysis problem falls right into the category of the off-policy variant of PG methods. In this method, the agent learns (trains and evaluates) solely from historical data, without online interaction with the environment. Thus, the transition probabilities are not considered in this algorithm, and we have not considered them either in our approach.

The policy network can robustly estimate the selection probability surface that we call the behavioral policy, denoted as π_θ , where θ is the vector of parameters of the network producing the behavioral policy. Now we aim to use the off-policy PG algorithm to tune the network parameters to produce the optimal selection probability surface denoted as π_{θ^*} , where θ^* is the vector of parameters of the tuned network producing the optimal selection probability surface given the current state. The gradients of the network tell us how the network should modify the parameters if we want to encourage any decision (action) in the future. We modulate the loss for each action taken in a possession according to their eventual outcome, since we aim to increase the probability of successful actions (with higher rewards) and decrease it for the unsuccessful ones. We train the policy network

with the help of the gradient vector, which encourages the network to slightly increase the likelihood of actions yielding large positive rewards and decrease the likelihood of negative ones. In the off-policy PG algorithm, the gradient can be defined as in (2):

$$\nabla_{\theta'} EPO(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=1}^H \nabla_{\theta'} \log \pi_{\theta'}(a_t | s_t) \left(\prod_{t=1}^H \frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)} \right) \left(\sum_{t=1}^H r(s_t, a_t) \right) \right] \quad (2)$$

where, θ is vector of parameters of the behavioral policy network (yielding behavioral probability surfaces), and θ' is the vector of parameters of the modified policy network that must be trained to yield optimal probability surfaces. Gradient vector $\nabla_{\theta'} EPO(\theta)$ is the gradient that computes a direction in the parameter space leading to modification in the probability of selection assigned to each location on the field. Consequently, actions with a high reward will tug on the probability density stronger than actions with a low reward. Thus, the off-policy PG algorithm helps the probability density to shift around in the direction of high rewarding actions, making them more likely to occur. The pseudo-code for PG algorithm is provided in Figure 12 in Appendix A.1.

5. Evaluating the outcome of the optimal decisions prior to deployment in a match

In the RL domain, evaluating a policy (decision) concisely means summing the rewards yielded by the selected actions (ball destination) given policy π . In our soccer analytics task, this type of evaluation would not be practical, as we cannot change the actions that players performed in the past (e.g., change the pass direction in a given situation) and see what would have been the outcome of the possession. In order to provide an intuitive evaluation for the outcome of our derived optimal policy to the soccer coaches and other decision makers prior to deployment in a real match, we compare the performance of the behavioral and optimal policies through three different evaluation methods: importance sampling, doubly robust, and through looking at the outcome of the reconstructed possessions from output 3 of the policy network (see figure 4).

5.1. Off-policy policy evaluation

After applying the off-policy PG algorithm to obtain the optimal policy in our soccer analysis task, we face the following challenge: while training can be performed solely from historical actions on the pitch, the evaluation unfortunately cannot, because deploying the optimal policy in a real soccer match would be too expensive to test its performance. This challenge motivated us to use off-policy policy evaluation (OPE), which is a technique for testing the performance of a new policy, when the real-world deployment is near to impossible due to being expensive or time-consuming. With OPE, we estimate the performance of our optimal policy based on the historical match dataset of the actions performed by players and possibly dictated by the coaches. To do so, we test the policies with two methods: importance sampling and doubly robust. Both methods take samples from the behavioral policy π_{θ} (i.e., actions performed by the players in the past) to evaluate the performance of optimal policy $\pi_{\theta'}$. Please see the blog post in [20] for the concepts of OPE methods, and research papers [17,18,19] that proposed these types of evaluations.

5.2. Evaluating the outcome of reconstructed possessions with new policy

Another option to evaluate the optimal policy is to reconstruct sequences of actions via the optimal policy network. Note that our designed policy network has a surprisingly good capability of reconstructing the actions through the decoder part in output 3 (see Figure 4). Before applying PG to the policy network, the decoder part could reconstruct the sequences of actions following from the behavioral policy for each team. After applying PG to the policy network, the decoder part can reconstruct the actions following the optimal policy, and we can assign rewards to each of the reconstructed actions, summing the yielded rewards, and evaluate the optimal policy.

5.3. Evaluating expected possession outcome of the behavioral and optimal decisions

With the proposed evaluation techniques (i.e., importance sampling (IS), doubly robust (DR), and reconstructed actions (Sim)) we proceed with evaluating the derived policies in terms of deciding the ball destination. Figure 7 illustrates the Kernel Density Estimation (KDE) of EPO yielded following the behavioral policy using (1) and following the derived optimal policy using the techniques described in the previous sections. Note that we only trained one optimal policy and evaluate it with three different methods. Moreover, the method that estimates larger EPO is not necessarily the best off-policy policy evaluator. The proposed evaluation methods are merely used to compare the behavioral policy with the obtained optimal policy. We can see in Figure 7 that the density of EPO in all possessions of the 2020-2021 season of the Belgian Pro League is focused around -0.01 following the behavioral policy (blue curve), meaning that on average, a possession in this league will lead to a goal against with 1% chance. However, if the teams follow the optimal policy, this number increases to better outcomes. For instance, the IS (orange curve) predicts that if the teams follow the optimal policy, the EPO in all possessions will be focused around 0.015, meaning that on average, a possession in this league will lead to a goal with 1.5% chance.

More intuitively, given an average of 100 possessions per team in a game, they would improve their expected goal difference per game by 2.5 goals, from -1 to +1.5 evaluated by IS, if they changed their behavior to the optimal policy. The other evaluators (DR and Sim) exhibit similar improvements as well.

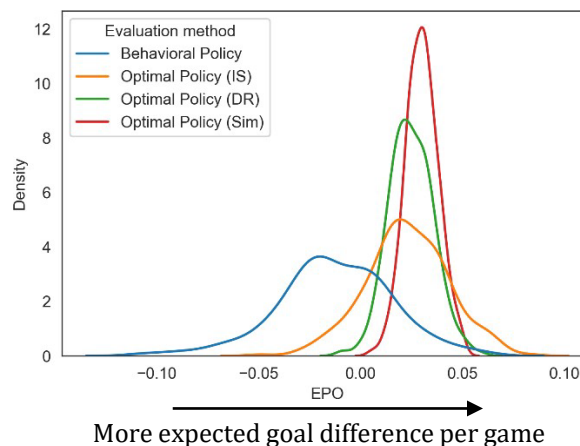


Figure 7. KDE comparing the league-wide model's EPO derived following each of the optimal and behavioral policies on 330 games of 2020-21 Belgian Pro League. The behavioral policy yields -0.01 EPO (likely to lose the game by 1 goal), whereas the optimal policy yields higher expected goal difference predicted by all evaluators.

6. Use cases

While our proposed approach captures a fine-grained analysis of the behavioral policy and proposes the optimal policy in the continuous action space (each location on the pitch), soccer practitioners such as coaches and opposition analysts might be more interested in seeing how their team has performed in different zones on the pitch and if they could modify their strategy to maximize their chances of winning their upcoming games. To this end, we discretize the action space with regards to different zones on the pitch and different possession phases to analyze the teams in the Belgian Pro League.

In this section, we first analyze the momentary optimal and behavioral decisions on the full surface of the pitch, depicting two examples of successful and unsuccessful possessions. Second, we distinguish the teams in terms of their behavioral and optimal policies and the improvement in their EPO and expected goal difference. Third, we partition the pitch into five zones: (left and right wings, left and right half-spaces, centre), where the propensity of the pass directions (forward, backward, sideways) and shot in the league-wide and team-specific models can be analyzed. Fourth, we analyze the action propensities in different possession phases (build-up, established possession, attack), where we show how often teams should perform short passes, long balls and shots to maximize their chances of winning.

6.1. Momentary decision-making analysis on the full surface of the pitch

The derived optimal policy can help players and coaches evaluate the outcome of a possession at any point in a game in the following two ways:

- Evaluate how the actions that the team performed (i.e., behavioral policy) contributed to the evolution of their EPO throughout the game.
- Compare the outcome of the actions that the team performed with the actions that the optimal policy suggests, and how those optimal actions would have contributed to the evolution of their EPO throughout the game.

In order to demonstrate a momentary situation in a game, we analyze two different possessions in the game between Cercle Brugge and Union SG in the 2021-22 season of the Belgian Pro League, which Union SG won with 0-3. Newly promoted Union SG has performed extremely well in the first half of the 2021-2022 season with the club topping the league table at the time of writing this paper.

1. **A successful Union possession:** Figure 8 shows the sequence of frames for the 59:26 – 59:32 time window that resulted in a goal for Union. The Union players are shown as green circles. The first row shows the selection surface of their behavioral policy, learned from their previous matches in Belgian Pro League 2021-22. The second row shows the immediate reward they would earn for moving the ball to each pitch location. At the first frame, Undav possesses the ball and passes backward to Teuma (T1:27s), adding about 0.003 to EPO. Then Teuma successfully finds the optimal location of the pitch by forward passing the ball to Undav in center (T2:29s). Then Undav immediately performs sideways pass to Vanzeir (T3:30s). Finally, Vanzeir performs successful shot (T4:32s), culminating the EPO curve. Below the frame-by-frame sequence, the EPO evolution is shown for Union SG, according to both the behavioral policy and the optimal policy. Union's behavioral policy is quite close to the optimal one. The relatively small EPO difference of 0.00031 (i.e., the difference between

the EPO's resulting from the optimal and behavioral policies) shows their near-optimal performance during this possession. (Link to the video of this possession ⁴)

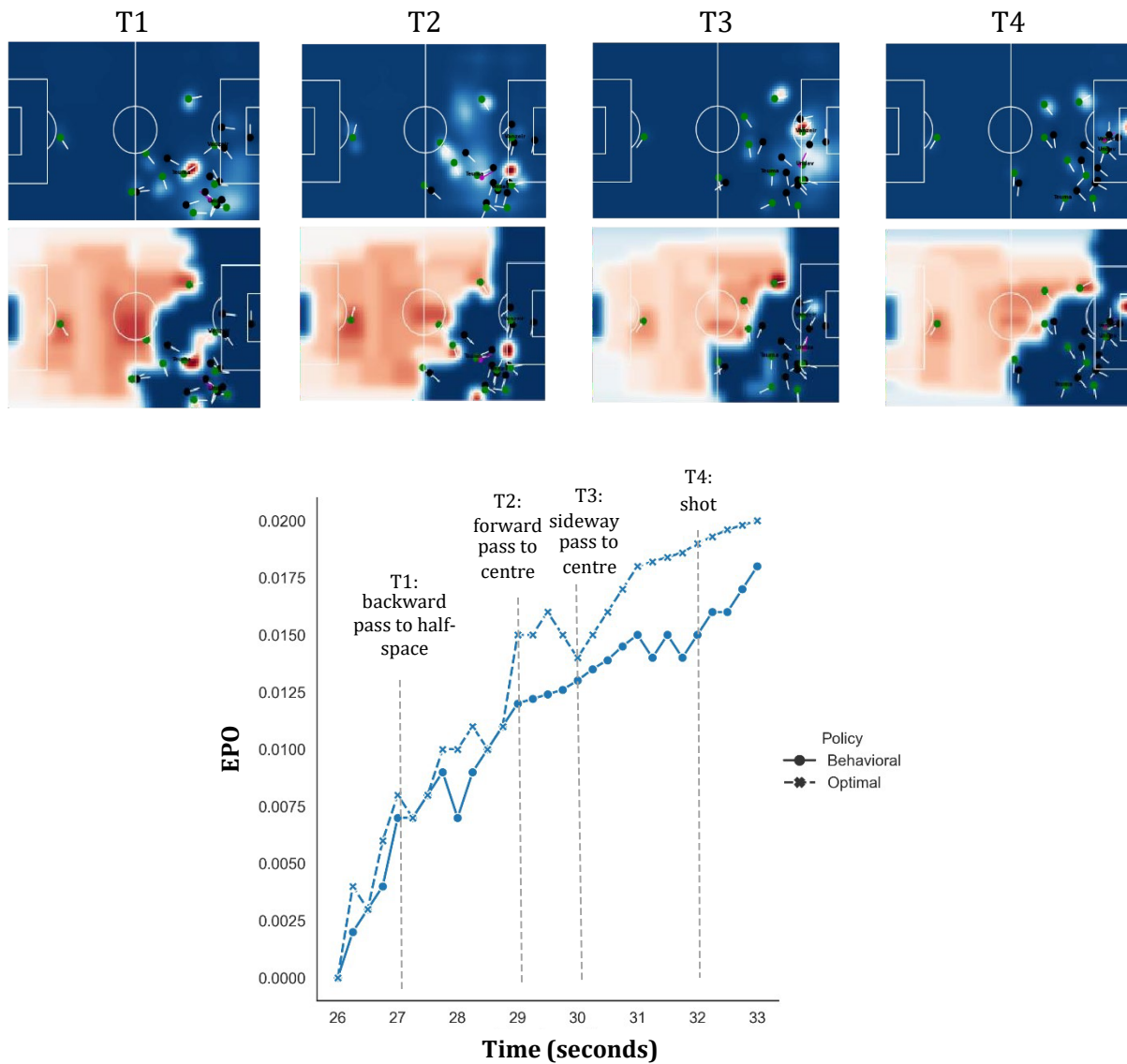


Figure 8. A successful possession for Union in their game against Cercle Brugge in the 2021-22 season of the Belgian Pro League. The Union players, who are shown as green circles, are in possession of the ball. The top row shows the selection probability surfaces according to their behavioral policy. The middle row shows the immediate rewards they would have earned if they had moved the ball to a certain pitch location. The bottom row shows the EPO curve according to the behavioral and optimal policies.

2. **An unsuccessful Cercle Brugge possession:** Figure 1 shows the sequence of frames for the 24:13 – 24:23 time window that did not result in a goal for Cercle Brugge. The figure again

⁴ bit.ly/3rzJefj

shows the behavioral selection surfaces, reward surfaces, and EPO curves. The Cercle Brugge players are shown as green circles. In the first frame, Daland possesses the ball (T1:14s) and dribbles. Between seconds 15 and 16, the optimal policy suggests passing the ball to his teammate Kanoute (i.e., the direction of the green arrow in the left surface). However, Daland decides to continue dribbling until (T:19s). The EPO curve shows that the EPO would have increased if Daland had followed the optimal policy. In the remainder of the possession, Daland passes the ball to Hotic (T: 19s), which is an optimal action and both EPO curves are improving. Later, Hotic continues dribbling until he performs an unsuccessful attempt (T:23s) (i.e., the direction of the red arrow in the right surface) that results in a significant decline of the EPO curve to -0.005, while the optimal policy suggests a forward pass from Hotic to Matondo at (T2:20s) (i.e., the direction of the green arrow in the right surface), so Matondo would have a higher chance of scoring according to the estimated xG at his point. That is the reason for the significant divergence between the EPO for the optimal policy and the behavioral policy from (T2:20s). The EPO difference for Cercle Brugge for this possession is 0.014, which is quite large and demonstrates the bad decisions that the Cercle Brugge players took during this possession. (Link to the video ⁵).

6.2. Team-specific improvements after adopting the optimal strategies

We showed the behavioral and optimal action propensities for a typical team in the Belgian Pro League. However, such propensities are likely to vary according to the different team's playing styles. While our league-wide model analyzes and optimizes the expected behavior of a typical team in the league, coaches are interested in team-specific details that set them apart from their opponents. To this end, we provide a fine-grained analysis of the behaviors for each specific team in the league and provide the optimal decisions as well as the number of potential additional goals they could score throughout a game. To quantify the effect of changing their policy, we use a metric called EPO difference that is interpreted as the increase in likelihood of a possession ending in a goal, if the team changed their behavioral policy to the optimal one: $EPO\ difference = EPO(\theta^*) - EPO(\theta)$. Considering 100 possessions for each team per game, the improvement in expected goal difference is calculated as: $EPO\ difference \times 100$.



















In this section, we evaluate the performance of following the derived optimal policy specified for each team in the Belgian Pro League. We develop a league-wide model by training the network on all games of the 2020-21 season, and then fine-tune the policy network to develop a team-specific model for each of the 18 teams using the games of the 2021-22 season until the moment of writing this paper (October 2021). What distinguishes our proposed method from the state-of-the-art is that we can directly evaluate the performance of following the optimal policy rather than performing counterfactual reasoning and analyzing what-if scenarios (e.g., What could have happened if the team increased their long or short pass/shot frequency by X%?).

We explore the effect of changing the policies for the teams from their behavioral policy to the optimal policy in terms of EPO and addition to the expected goal difference. For each team, we obtain the mean EPO over all possessions in the 2021-22 season when following their behavioral policy and when following the optimal policy. Table 2 presents the calculated differences. The teams are sorted according to the league table at the time of writing this paper. For instance, the mean EPO over all

⁵ bit.ly/3HC2j5X

possessions of Club Brugge in the 2021-22 season is calculated as 0.0046, meaning 0.46 goals per game. We used the off-policy evaluation (importance sampling) to evaluate the EPO of their possessions if they had followed our derived optimal policy. The results show that Club Brugge could have increased their likelihood of ending possessions in a goal by 0.016, by following our proposed optimization. So, they would have improved their expected goal difference by 1.6 goals per game. Another observation from Table 2 is that the optimal policy yields a smaller improvement in EPO and goal difference for the teams at the top of the table, and a larger improvement for the teams at the bottom of the table. That is because a team like Union SG at top of the table is quite often selecting the optimal actions (their behavioral policy is nearly the same as their optimal policy), whereas the behavioral policy of the teams at the bottom of the table is far from their respective optimal policy.

Table 2. The effect of changing the policy on the team-level from the behavioral to the optimal policy.

Teams	Mean EPO (behavioral)	Mean EPO (optimal)	EPO difference	Improvement in the expected goal difference per game
 Union SG	0.0160	0.0210	0.005	0.5
 Club Brugge	0.0046	0.0215	0.016	1.6
 Antwerp	0.0053	0.0153	0.010	1.0
 Mechelen	0.0010	0.0130	0.012	1.2
 Charleroi	0.0040	0.0180	0.014	1.4
 Anderlecht	0.0060	0.0170	0.011	1.1
 Genk	0.0020	0.0230	0.021	2.1
 Eupen	0.0013	0.0123	0.011	1.1
 Kortrijk	0.0001	0.0181	0.018	1.8
 Gent	0.0046	0.0246	0.020	2.0
 Oostende	-0.0086	0.0064	0.015	1.5
 St Liège	-0.0040	0.0150	0.019	1.9
 Sint-Truiden	-0.0040	0.0200	0.024	2.4
 OH Leuven	0.0027	0.0287	0.026	2.6
 Waregem	-0.0080	0.0170	0.025	2.5
 Seraing	-0.0046	0.0244	0.029	2.9
 Cercle Brugge	-0.0040	0.0240	0.028	2.8
 Beerschot	-0.0087	0.0233	0.032	3.2

6.3. League-wide model for optimal action selection in different zones

In the last decades, many soccer experts such as Guardiola and Klopp have focused on a zone between the center and the wings, which are typically called the half-spaces, in their game strategy and tactics. Although the center is closest to the goal and the players can pass in any direction (backward, forward, sideways) or shoot, this zone is very crowded in most situations. Furthermore, the wings are less crowded, but they are further from the goal. In contrast, in the half-spaces, the ball is close enough to the goal and players have enough space to pass in any direction. This led us to analyze the pass and shot propensities in each of the zones.

Figure 9 illustrates the pass direction and shot statistics in different zones of the pitch (shown at the left side) for 18 teams in the Belgian Pro League. Our proposed method first analyzes the behavioral propensities of the teams to choose a backward pass, a forward pass, a sideways pass or a shot in each of the zones on the pitch, and then derives the optimal propensities. The results show that in a league-wide model, teams could increase their Expected Possession Outcome (EPO) by 0.025, meaning that they could increase the likelihood of a possession ending in a goal by 0.025 if they had adopted the optimal policy. Considering a team with 100 possessions per game on average, this improvement would lead to +2.5 expected goal difference per game. Although such an improvement is not much realistic in the real-world as it assumes that all players can perfectly follow the optimal policy (which is not the real case), it can give an interpretable insight to the coaches and analysts on adjusting their current pass and shot propensities. Table 3 provides the optimization results for the Belgian league-wide model. The numbers in the table represent the percentage of increase or decrease in action propensities if the teams adopted the optimal policy. Overall, the optimization results propose a larger propensity of backward passes on the wings, a smaller propensity of backward passes and a larger propensity of shots in the half-spaces, and a larger propensity of forward passes and sideways passes in the center. The result of the league-wide model provides an insight to the coaches and analysts to capture the performance and the amount of adjustment in the action tendency of an ordinary team in the Belgian Pro League playing with any other team in or out of the league. In the following sections, we provide the optimization results for any specific team that could help them to beat the opponents and climb the league table.

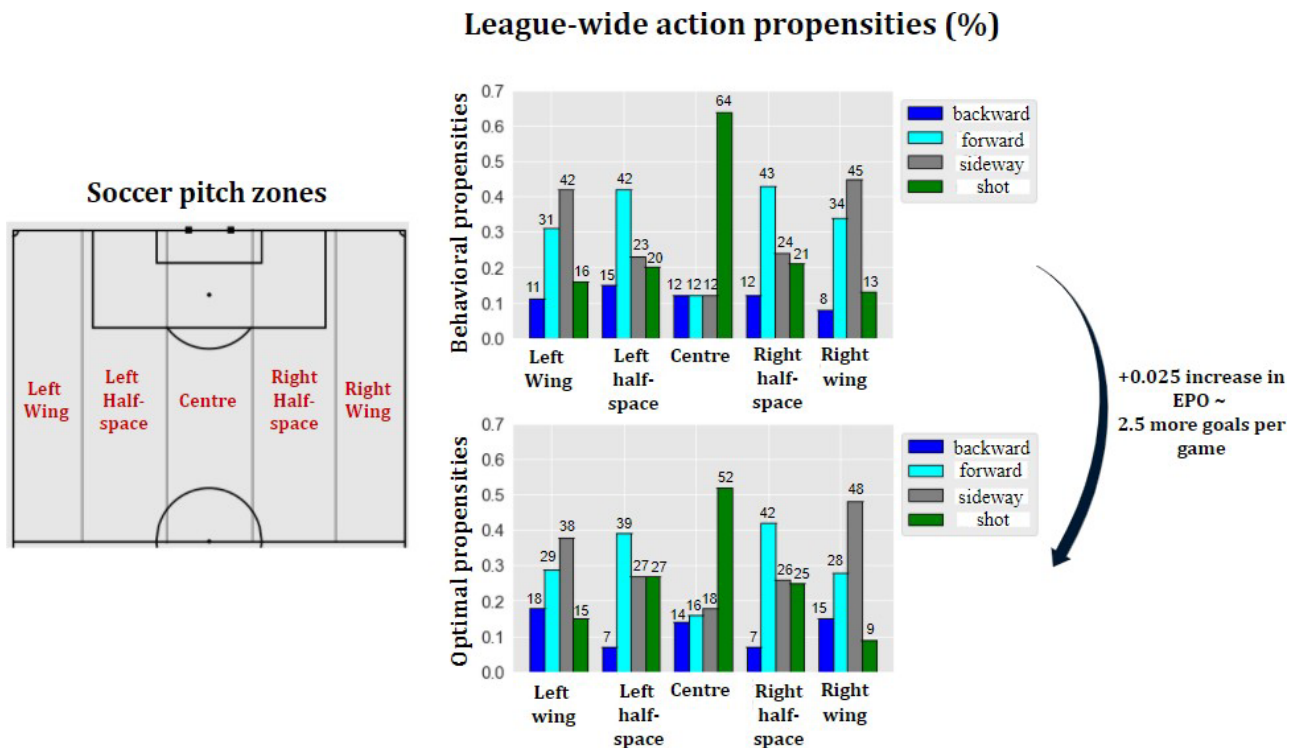


Figure 9. League-wide analysis of the Belgian Pro League 2020-21 season in terms of team propensities to perform backward passes, forward passes, sideways passes and shots in different zones of the pitch. Our optimization model proposes specific modifications to the behavioral propensities and would lead the teams to improve their expected goal different per game by 2.5 goals.

Table 3. Percentage change in adopting each of the four actions in each of the five zones according to the Belgian league-wide model. The + and – signs represent the relative increase and decrease in action propensities of the behavioral policy compared to the optimal policy.

	Left wing	Left half-space	Centre	Right half-space	Right wing
Backward	+63%	-53%	+16%	-41%	+87%
Forward	-6%	-7%	+33%	-2%	-17%
Sideway	-9%	+17%	+50%	+8%	+6%
Shot	-6%	+35%	-18%	+19%	-30%

6.4. Where should the teams pass or shoot in different zones?

In this section, we customize the use case from Section 6.1 on the team level to provide the intuitive optimal suggestions for the coaches and players of specific teams in the league. We show the result for two competing teams with different playing styles: Club Brugge at the top, and Cercle Brugge at the bottom of the league table until October 2021. Following the behavioral policy, Club Brugge is likely to score 0.46 goals per game, while Cercle Brugge is likely to concede 0.40 goals per game. Our proposed optimization method is likely to yield 1.6 improvement in expected goal difference for Club Brugge, and 2.8 improvement in expected goal difference for Cercle Brugge, if they adopted their policies according to Figure 10. In summary, the optimal playing style for Club Brugge is to increase the number of backward passes on the wings and the number of forward passes and shots in the half-spaces, whereas the optimal playing style for Cercle Brugge is to increase the number of sideway passes on the wings, the number of forward passes in the half-spaces and the number of shots from the center.

Optimal playing style in different pitch zones

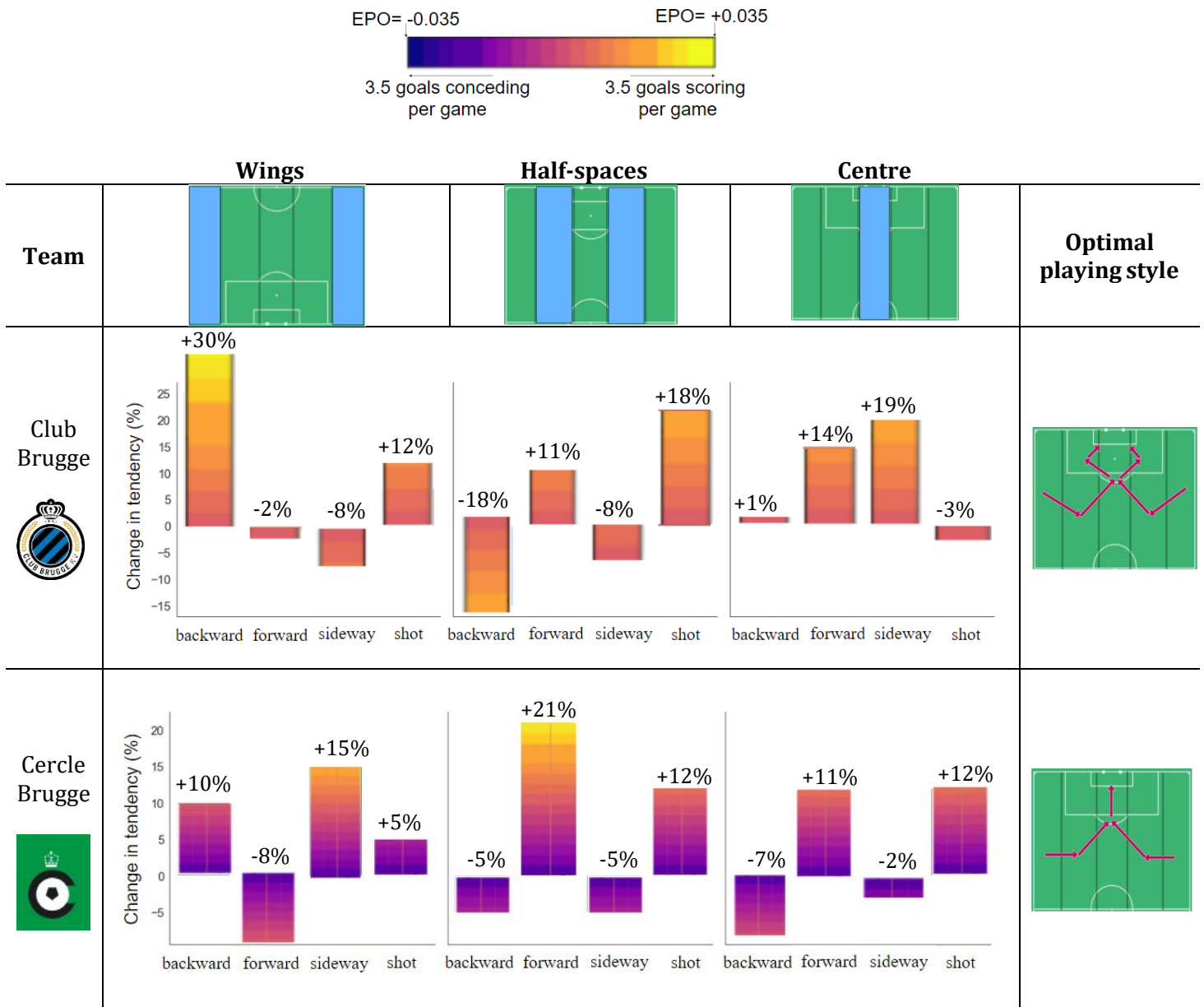


Figure 10. Optimal modifications in action tendencies with respect to different zones for Club Brugge and Cercle Brugge. The color in the bar charts represents the amount of EPO the teams can gain with respect to the percentage of modifications in their action tendencies.

6.5. Long vs short distance passes and shots in different possession phases

Besides providing behavioral and optimal propensities in different pitch zones, our framework allows to analyze the propensities of playing short passes, playing long balls and shooting in each of the different possession phases. To do so, we use an action space consisting of three actions, namely “short pass”, “long ball: ball move over than 32 meter in length” and “shot”, and the possession phases defined in Section 4.1 (i.e., build-up, established possession, attack). Figure 11 illustrates the behavioral and optimal propensities as well as the EPO difference for the four teams in the Belgian Pro League with different playing styles. In summary, the optimal policy proposes:

- **Club Brugge:** This team’s propensity of playing short passes in the different phases is close to optimal. The optimal policy advises them more long-distance shots, and to play fewer long balls from short distance (e.g., attack phase).
- **Oostende:** The optimal policy advises them to attempt more long-distance shots, and more short passes and long balls from short distance (e.g., attack phase).
- **OH Leuven:** This team’s propensity to attempt short-distance shots is close to optimal. The optimal policy advises them to play more short passes in the build-up and established possession phases.
- **Zulte Waregem:** This team’s propensity to playing long balls is close to optimal in all phases. The optimal policy advises them to play more short passes in build-up, and to attempt more shots in the attack phase.

The analysis for the remaining teams in the Belgian Pro League is provided in Figure 13 in Appendix A.2.

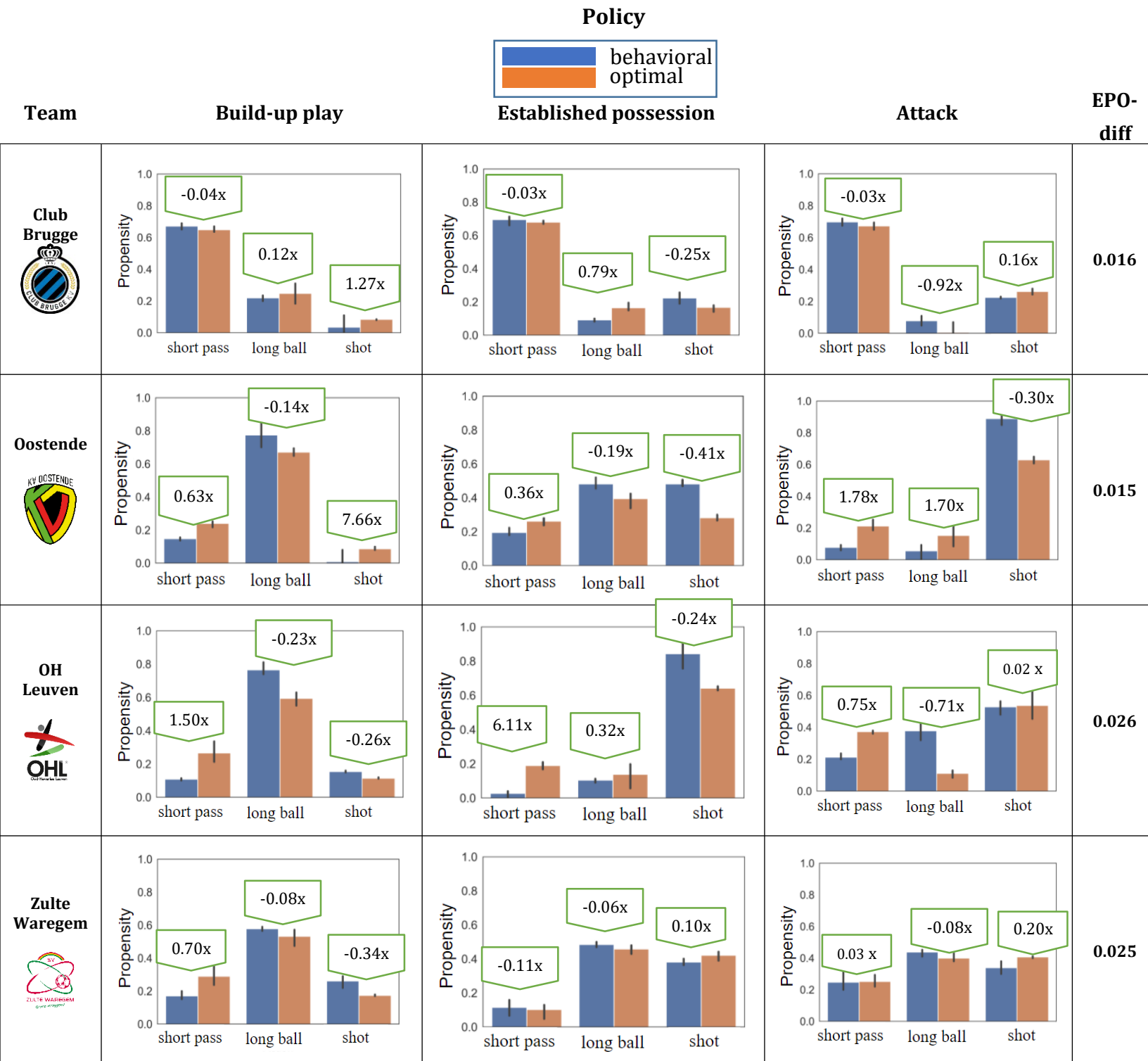


Figure 11. Team-specific propensities to play short passes, to play long balls and to attempt shots in three different possession phases.

7. Related work

The state-of-the-art methods in soccer analytics are categorized into decision making analysis, and action valuation methods. Current analytics methods on soccer decision making are limited to measuring the potential outcome of the alternative decisions (e.g., What would have happened if the team had increased particular action frequency by x%?) rather than directly discovering the optimal ones. Exploring the effect of changing such decisions is studied [6,7,16] in football, and [21,22] in basketball. A comprehensive method must consider a wide set of actions and all exact player- and ball locations rather than team formations. The current methods using artificial intelligence could predict where a player will pass the ball (pass selection) [10], the likelihood of that pass being completed (pass success) [9,10], and whether this pass will result in a scoring opportunity (pass valuation) [1,2,3,4,5,6]. These latter values are estimated by the probability of a shot being made in the next 10 seconds, next 10 actions, or at the end of the possession. Our work uses deep learning techniques to analyze the current strategy of the teams in terms of their propensities to select the destination location of the ball on the full pitch surface, and then use RL to directly discover optimal tactics based on all players and the ball positions on the pitch.

8. Conclusion

We propose an end-to-end deep reinforcement learning framework that derives optimal decisions solely from the teams' actual behaviors. To do so, we first analyze the current strategy of the teams in terms of their propensities to select the destination location of the ball using deep learning, and then discover the optimal tactics based on all players' locations on the pitch. Furthermore, our optimization framework continuously highlights the optimal space on the pitch for the ball to be moved to, which has the maximum potential contribution to the team winning in the long-term, even if a particular on-ball action does not directly contribute to a goal. We directly derive the optimal policy for teams in terms of moving the ball instead of evaluating the effect of alternative policies and counterfactual reasoning such as: What would have happened if another ball destination had been selected? In addition to continuously highlighting the optimal ball destination, we compare the team-specific behavior in terms of selecting each of the discrete actions (i.e., backward pass, forward pass, sideway pass, shot) with the optimal ones and show for each team the number of additional goals they would score if they adapted their current behavior to the optimal behavior. Concretely, in the league-wide model, we show that teams would improve their expected goal difference by +2.5 goals from 1 goal against to 1.5 goals for, if they would adapt their current strategies to the optimal ones. The applications of our approach for soccer coaches and players are multifold: First, analyzing their actual action selection propensities in each game context. Second, obtaining a team-specific optimal behavior for each game situation. Third, measuring how much their current strategy differs from the optimal one. Fourth, estimating the expected improvement in terms of goal difference in future games if they adopted the optimal strategy. Last, finding the optimal ball destination for each game situation. To the best of our knowledge, our work is the first attempt to use reinforcement learning to maximize the expected possession outcome for the full pitch surface in soccer. A direction for future work is to expand the current framework to capture the contributions of individual players such that they could adapt themselves to their team-specific optimal strategy.

Acknowledgement

Project no. 128233 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the FK_18 funding scheme. We thank Maaïke Van Roy for her valuable comments that helped improve this paper.

References

- [1] G. Liu and O. Schulte, "Deep reinforcement learning in ice hockey for context-aware player evaluation," in International Joint Conference on Artificial Intelligence, 2018.
- [2] G. Liu, Y. Luo, O. Schulte, and T. Kharra, "Deep soccer analytics: learning an action-value function for evaluating soccer players," Data Mining and Knowledge Discovery, vol. 34, no. 2, 2020.
- [3] K. Sing, "Introducing expected threat (xt) modelling team behaviour in possession to gain a deeper understanding of buildup play." 2018. [Online]: <https://karun.in/blog/expected-threat.html>
- [4] L. Gyarmati and R. Stanojevic, "Qpass: a merit-based evaluation of soccer passes," in KDD Workshop on Large-Scale Sports Analytics, 2016.
- [5] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer." in ACM KDD, 2019.
- [6] J. Fernandez, L. Bornn, and D. Cervone, "Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer," in MIT Sloan Sports Analytics Conference, 2019.
- [7] Maaïke Van Roy, Pieter Robberechts, Wen-Chi Yang, Luc De Raedt, Jesse Davis. Leaving Goals on the Pitch: Evaluating Decision Making in Soccer. in MIT Sloan Sports Analytics Conference, 2021.
- [8] Bransen, L., Robberechts, P., Van Haaren, J., & Davis, J. Choke or Shine? Quantifying Soccer Players' Abilities to Perform Under Mental Pressure. in MIT Sloan Sports Analytics Conference, 2019.
- [9] Power, P., Ruiz, H., Wei, X., Lucey, P.: Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. in ACM KDD, 2017.
- [10] J. Fernandez and L. Born, "Soccermap: A deep learning architecture for visually-interpretable analysis in soccer," in ECML PKDD, 2020.
- [11] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, Iain Matthews. "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. in MIT Sloan Sports Analytics Conference, 2015.
- [12] Filippo Studzinski Perotto, Laurent Vercouter. Tuning the Discount Factor in Order to Reach Average Optimality on Deterministic MDPs. International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2018.
- [13] Sergey Levine, Vladlen Koltun, "Guided policy search: deep RL with importance sampled policy gradient", Proceedings of the 30th International Conference on Machine Learning, 2013.
- [14] Schulman, L., Moritz, Jordan, Abbeel. Trust region policy optimization: deep RL with natural policy gradient and adaptive step size, 2015.
- [15] Schulman, Wolski, Dhariwal, Radford, Klimov. Proximal policy optimization algorithms: deep RL with importance sampled policy gradient, 2017.
- [16] J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect, 1st ed. USA: Basic Books, Inc., 2018.
- [17] T. Xie, Y. Ma, and Y. Wang, "Optimal off-policy evaluation for reinforcement learning with marginalized importance sampling," CoRR, 2019.

- [18] M. Farajtabar, Y. Chow, and M. Ghavamzadeh, "More robust doubly robust off-policy evaluation," CoRR, 2018.
- [19] J. Nan and L. Lihong, "Doubly robust off-policy evaluation for reinforcement learning," CoRR, 2015.
- [20] <https://ambiata.com/blog/2020-10-27-off-policy-evaluation/>
- [21] N. Sandholtz and L. Bornn, "Replaying the NBA." In Proceedings of the MIT Sloan Sports Analytics Conference. 2018.
- [22] N. Sandholtz and L. Bornn, "Markov Decision Processes with Dynamic Transition Probabilities: An Analysis of Shooting Strategies in Basketball.", 2020.

Appendix

A.1. Sample pseudo-code for training optimal policy network using Policy Gradient

Algorithm 1 Policy Gradient

```

1: Initialize  $\theta$  from the behavioral policy of the teams
2: for each team in the league do
3:   for each game of a team in the league do
4:     for each possession  $\tau_i \sim \pi_\theta$  in the game with the performed actions and earned rewards do
5:       for  $t=1, \dots, H-1$  do
6:          $\theta \leftarrow \theta + \alpha \Delta_\theta EPO(\theta)$        $\triangleright$  Update the parameter of the network to increase the probability of highly
           rewarding actions
7:       end for
8:     end for
9:   end for
10: end for
11: return  $\theta$ 

```

Figure 12. pseudo-code for training optimal policy network to obtain optimal probability surfaces

A.2. Team-specific behavioral and optimal action propensities for all teams

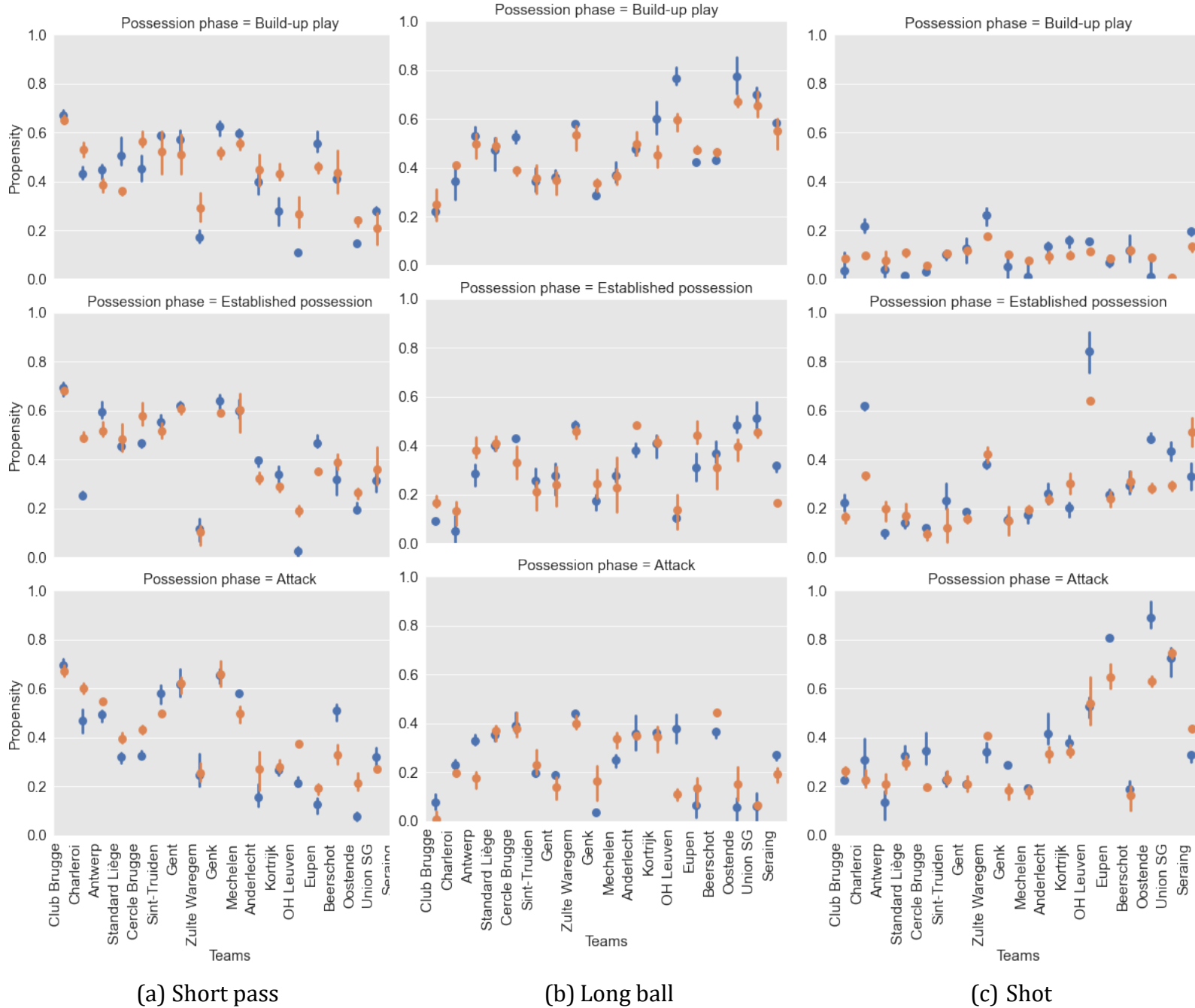


Figure 13. Team-specific action propensities with respect to the different phases of possession. Blue color represents behavioral propensity, and orange color represents optimal propensities. The circles and vertical lines represent mean and standard deviation of the propensities over all possessions of the teams in our dataset. The vertically closer of two circles for each team represents the more optimal performance of the team for each action type with respect to the possession phase.