# SPATIOTEMPORAL SIMULATION OF SOCCER MATCH EVENTS

by

Oden Petersen (z5220271), Bodu Gong (z5320212), Laeeque Jamdar (z5218707)

TABLE OF CONTENTS

LIST OF TABLES . . . . . iii

LIST OF FIGURES . . . . . iv

ABSTRACT . . . . . 1

BRIEF OVERVIEW OF THE GAME OF SOCCER . . . . . 2
Subject Matter Considerations for Modeling . . . . . 3
Goal Rarity . . . . . 3
Markov Assumption . . . . . 4
Waiting Times . . . . . 5

DATASET . . . . . 7

METHODOLOGY . . . . . 7
Pre-Game Prediction . . . . . 8
Logistic regression . . . . . 8
Direct Prediction of Future Goals . . . . . 10
Modeling Expected Waiting Time . . . . . 10
Modeling Event Details . . . . . 11
Deducing Conditional Waiting Time Distributions . . . . . 11
Synthesising an Outcome Distribution from Conditional Waiting Times . . . . . 12
Recursive Probabilistic Simulation . . . . . 13
Handling Numerical Features . . . . . 14

EVALUATION . . . . . 16

DISCUSSION . . . . . 17
Limitations of the Method . . . . . 17
Errors in Parameter Estimation . . . . . 17
Applications of the Stochastic Simulation Framework . . . . . 17
Applications to Team Management . . . . . 17

APPENDIX . . . . . 18

REFERENCES . . . . . 19

LIST OF TABLES

Table 1. Event frequencies in the England dataset . . . . . 18

LIST OF FIGURES

Figure 1. Frequency heatmap of the number of goals scored by both teams across all matches in the England dataset, with white being highest frequency and black being lowest . . . . . 3
Figure 2. Histograms of Waiting Times Prior to Various Event Types vs. PDF of Exponential Random Variable (both logarithmically transformed for easier visual comparison) . . . . . 6
Figure 3. Diagram Explaining the Kalman Filtering Technique . . . . . 14
Figure 4. Heatmaps of Field Position and Field Position Change . . . . . 15

ABSTRACT

Following the successful application of statistical methods to baseball analytics (sabermetrics) in the 1990s (see Lewis, 2010), analytics has become increasingly important to the strategic planning processes of professional sports teams, informing both team construction and player coaching. As the number of professional soccer games continue to grow around the world, it is imperative for coaches, team owners, and other stakeholders to use data driven insights to maintain their competitive advantage.

With the rise of electronic betting markets, including the recent introduction of blockchain-based prediction markets such as Augur and Polymarket, the established field of quantitative sports analytics is presented with a new opportunity for application as financial speculators seek to provide liquidity to retail bettors at intelligent prices via the use of automated market-making algorithms.

Furthermore, professional sports leagues have begun publishing live events data from matches for the explicit purpose of betting market pricing (see Pelit, 2022). Such data may also be synthesised from live match footage using computer vision techniques.

Beginning with the hypothesis that this live data contains meaningful information about match outcomes, this paper explores a variety of techniques for data-driven match outcome prediction. From the key assumption that the time series of match events satisfies the Markov property, we develop a baseline technique for predicting events of specific types (goals in particular), and then extend this into a more sophisticated and apparently novel simulation framework inspired by a convenient property of the exponential distribution.

Betting markets generally depend on either the exact scores of each team at the end of the game, or some function of these (including the special case of which team wins). We aim to construct a model that outputs a probability distribution over all ordered pairs of non-negative scores, as this is quite a general object that can then be easily processed into the desired outcome.

Finally, we discuss possible future applications of these models beyond the problem of live betting,

including to the problems of team construction, player coaching, and causal explanation of match events.

## BRIEF OVERVIEW OF THE GAME OF SOCCER

A game of soccer involves two opposing teams (typically a 'home' team and an 'away' team) of 11 players each, and is played on a rectangular field with two nets ("goals") set up on the short sides of the field. Each game consists of two identical 45-minute "halves" in which teams attempt to manoeuvre the ball into the goal belonging to the opposite team. At the end of the game, the team with the greater number of points is declared the winner (ties are dealt with separately).

Within these halves, there are a number of reasons play might stop:

- There are rules governing players' conduct on the field, and the referee may stop the game to reprimand players who break these rules (this is known as a "foul")
- If the ball leaves the playing field, the team to have touched the ball most recently is held responsible, and their opponent is given the opportunity to resume play by throwing the ball inwards from the edge of the field
- When a goal is scored, the ball is placed back in the centre of the field and the team against whom it was scored is given control of the ball

Although simple, this game requires expertise in timing, accuracy and strategic efficacy of actions such as ball passes. A variety of tactics (both strategic and athletic) all serve to individually give players and teams a slight edge over their opponents. Players make strategic use of their athletic abilities, the restrictions imposed by the rules of the game, and their predictions/observations of the actions of opponents and team members in order to react appropriately, to maintain physical control of the ball, and to defend, deceive, or collaborate in service of their objective.

## Subject Matter Considerations For Modeling

Considered as a prediction problem, it is important to note that this is both a time series problem and a spatial modeling problem (hence the use of "spatiotemporal" data).

### Goal Rarity

Goals are quite a rare event in the dataset used - only 0.96% of the events in the dataset are tagged as being a goal, and only 2374 goals appear in total (cf. 1). As visible in Figure 1, the largest number of goals scored by one team in a single match was 7 and the most common score was 1-1.

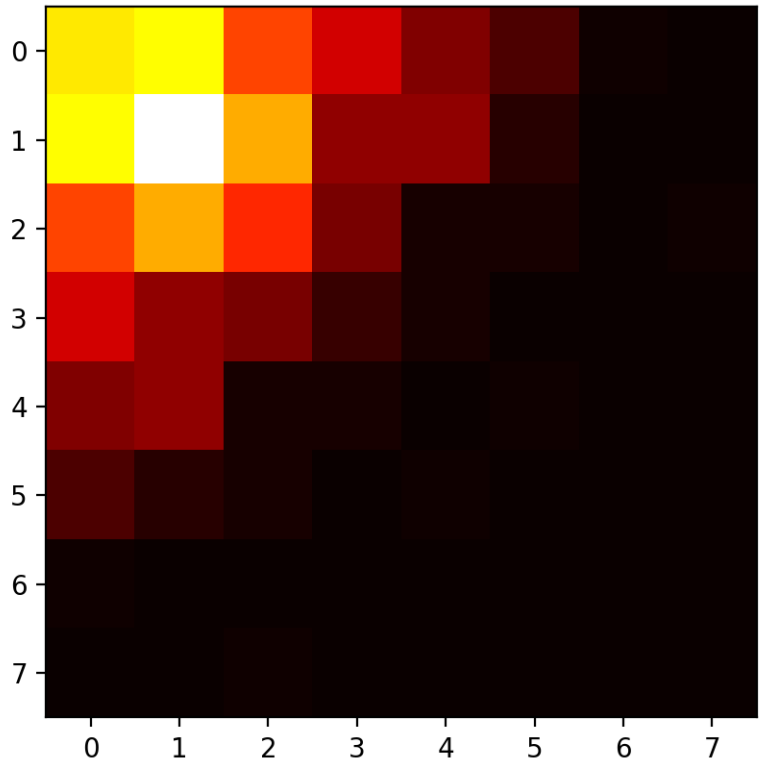


Figure 1. Frequency heatmap of the number of goals scored by both teams across all matches in the England dataset, with white being highest frequency and black being lowest

Extremely rare events create potentially pathological consequences for classical statistical estima-

tion techniques (e.g. see Taleb, 2020), because they greatly increase uncertainty (as measured by variance, differential entropy, or other metrics). While there are a large number of observed goals in the dataset, it is nonetheless important to apply caution and statistical rigour in analysing this problem. An analogy may be drawn to weather forecasting, which also attempts to predict the behaviour of a chaotic system prone to extremely rare but significant events, given both prior knowledge and present conditions.

Relatively high levels of uncertainty in model outputs would be a mildly disappointing result, but not hugely surprising in the special case of predicting goals far in advance of when they happen. However, an effective model should hopefully be able to predict goals with higher certainty close to when they occur, thus preventing informed bettors from exploiting an automated liquidity provision algorithm when goals are visibly imminent.

**Markov Assumption**

A stochastic process satisfies the Markov property if and only if the future behaviour of the system is statistically independent of all states of the system before the most recent state.

The ability of a team to score a goal requires that they have (or will gain) control of the ball and effectively manoeuvre it to a position from which the ball can be successfully launched into the goal. This will depend on factors such as the skill level of the individual players, the current ball position, and the current position of the players on the field.

Since none of these factors appears to be informed by match events prior to the most recent one, it seems reasonable to assume that the time series of match events is best predicted by looking only at the most recent recorded event.

One possible failure mode of this assumption is the contribution of different kinds of match events to the physical exhaustion of players. Ideally, well-trained professional players should not be greatly affected by this. Nonetheless, the events data might be augmented to resolve this non-Markovity by introducing a new feature attempting to track the cumulative exhaustion of players

(though we have not done this here). Such an augmentation would likely benefit from improved subject-matter knowledge about athlete exhaustion.

Throughout this analysis, we assume that the time series of match events satisfies the Markov property, and in special cases where this assumption is extremely inaccurate our modeling approach should be expected to underperform.

**Waiting Times**

After each event is observed, some unknown amount of time must pass before the next event is encountered. The typical choice for modeling these waiting times would be an exponential distribution, and the appropriateness of this distribution is confirmed empirically by visual comparison of the plots in Figure 1 (in general, the Kolmogorov-Smirnov test provides an algorithmic way of testing whether data follows a given distribution, though we have not used it here, as the histograms appear sufficiently suggestive of approximate exponentiality).

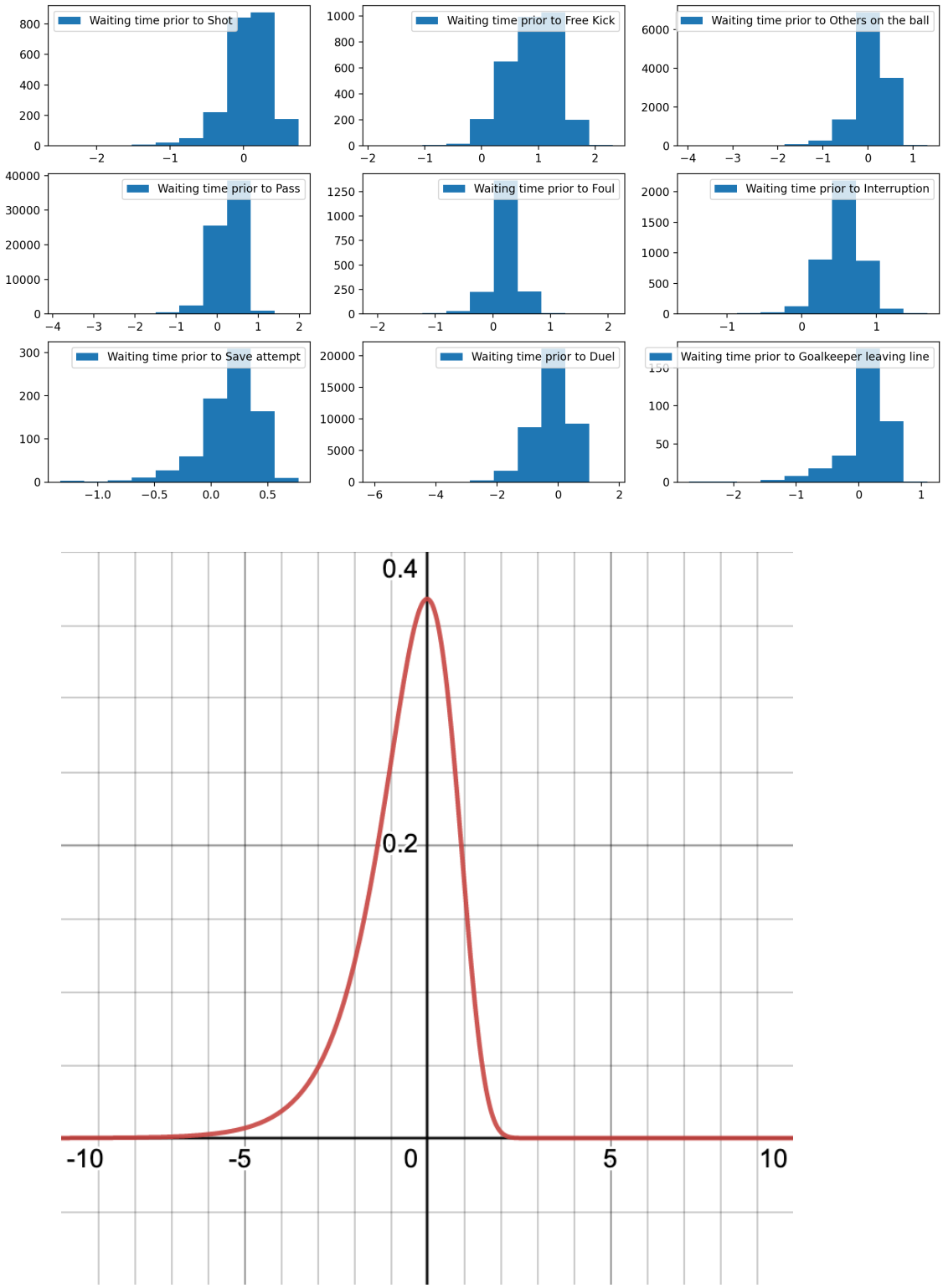


Figure 2. Histograms of Waiting Times Prior to Various Event Types vs. PDF of Exponential Random Variable (both logarithmically transformed for easier visual comparison)

The waiting time will typically be different leading up to different kinds of events; for example, if 20 seconds have passed already, it is much more likely that a free kick or free throw is being set up than that the players are preparing to pass. The significance of these conditional waiting time distributions is discussed further in the methodology section.

## DATASET

The dataset used in this project is described in detail in Pappalardo et. al., 2019. It consists of data about key events in professional soccer matches in various competitions, manually labeled by humans using specialised software, as well as data about the various matches, teams and players appearing in the table of events.

Throughout the project, we have restricted ourselves to using the England competitions data (extension of the methods to the entire dataset would be quite simple, but require more computational power).

The data has been extracted from Figshare using a publicly available Python API wrapper. Data from the various tables is then joined and parsed using a custom script (`parse.py`). The resulting CSV file is used for further analysis. A thorough amount of data cleaning was needed as most of the data consisted of jsons, strings and miscellaneous text inputs which were carefully removed before implemented in the predictive modelling stage.

## METHODOLOGY

In light of the motivations discussed for the problem, we are primarily interested in the supervised problem of predicting match outcomes in a principled manner that allows us to quote a probability distribution over possible ordered pairs of scores, given information about the teams, players, and most recent match event.

Techniques from the course are used where appropriate, and we make use of knowledge from probability theory to develop a novel framework for synthesising model outputs and producing

a probabilistic prediction. We begin with some baseline approaches, then add sophistication in search of improved accuracy.

To our knowledge, the synthesis of estimated waiting time and classification probabilities via properties of conditional exponential distributions as discussed below is not comparable to any published prior work (though it would not be surprising if something similar has been done before and we are simply unaware of it). It appears to be quite a general framework, and potential further applications (as well as some limitations) are briefly explored in the discussion section.

## Pre-game Prediction

Fair pre-match betting odds will depend on information about the teams, players and other match-specific details known prior to the beginning of the match.

In fact, one would hope that the prior distribution over possible match outcomes based solely on this information coincides approximately (e.g. as measured by Kullback-Leibler Divergence) with the initial estimate of the live prediction models prior to observing any events data.

Outcomes of matches prior to the train-test split date may also be of relevance, and may provide contextual information about the relative skill of different teams (e.g. the future outcome of A playing C is unlikely to be independent of the historical outcomes of A playing B and B playing C).

### Logistic Regression

We produce a prior prediction for each match by fitting a logistic regression to the outcome of previous matches in the current season.

We have one feature for each team, One-hot encoded so that the home team has a value of +1, and away team has a value of -1 and all the other teams have a value of 0. The target to predict is 1 if the home team won and 0 if they drew or lost.

To avoid rerunning the logistic regression for each match, we split the data into 6 chunks and use

the previous $k-1$ chunks to predict the matches in the $k$th chunk.

To tune the hyperparameter C for the logistic regression, we created a linear range in logarithm-space; that is our C-values were defined by `10**np.linspace(-2, 2,num=20)`. This is to effectively search a wide range. Table of results, percent_accuracy[C][chunk-1]:

```
0.01 [‘0.507’, ‘0.571’, ‘0.571’, ‘0.492’, ‘0.539’]
0.016 [‘0.507’, ‘0.571’, ‘0.571’, ‘0.492’, ‘0.571’]
0.026 [‘0.507’, ‘0.603’, ‘0.619’, ‘0.523’, ‘0.603’]
0.042 [‘0.507’, ‘0.650’, ‘0.666’, ‘0.650’, ‘0.603’]
0.069 [‘0.507’, ‘0.698’, ‘0.666’, ‘0.634’, ‘0.571’]
0.112 [‘0.571’, ‘0.730’, ‘0.666’, ‘0.634’, ‘0.523’]
0.183 [‘0.619’, ‘0.682’, ‘0.650’, ‘0.650’, ‘0.523’]
0.297 [‘0.619’, ‘0.730’, ‘0.650’, ‘0.650’, ‘0.507’]
0.483 [‘0.603’, ‘0.761’, ‘0.634’, ‘0.650’, ‘0.492’]
0.784 [‘0.682’, ‘0.761’, ‘0.634’, ‘0.634’, ‘0.492’]
1.274 [‘0.698’, ‘0.761’, ‘0.634’, ‘0.650’, ‘0.492’]
2.069 [‘0.698’, ‘0.761’, ‘0.634’, ‘0.650’, ‘0.492’]
3.359 [‘0.730’, ‘0.746’, ‘0.634’, ‘0.650’, ‘0.492’]
5.455 [‘0.730’, ‘0.746’, ‘0.634’, ‘0.650’, ‘0.492’]
8.858 [‘0.730’, ‘0.746’, ‘0.634’, ‘0.650’, ‘0.492’]
14.38 [‘0.698’, ‘0.730’, ‘0.634’, ‘0.650’, ‘0.492’]
23.35 [‘0.714’, ‘0.730’, ‘0.634’, ‘0.650’, ‘0.492’]
37.92 [‘0.698’, ‘0.730’, ‘0.634’, ‘0.650’, ‘0.492’]
61.58 [‘0.682’, ‘0.730’, ‘0.634’, ‘0.650’, ‘0.492’]
100.0 [‘0.682’, ‘0.730’, ‘0.634’, ‘0.650’, ‘0.492’]
```

## Direct Prediction of Future Goals

Combining the pre-game data with information about the most recent event (in accordance with the Markov property assumption), we can build separate models for:

- A probability distribution over the categorical information associated with the next event of interest; in particular, who the next goal will be scored by
- An estimate for the waiting time until the next event, conditional only on the most recent event

Using the outputs of these two models, we will show that a probability distribution for the final game outcome can be derived.

### Modeling Expected Waiting Time

Supervised regression techniques such as Linear Regression, Lasso and Ridge Regularised Regression, Decision Tree Regression and Gradient Boosted Regression (XGBoost) all allow for better-than-baseline estimation of expected waiting time.

We split matches into train and test sets based on whether they were before or after a given date. By engineering a variety of dummy variables and making use of hyperparameter optimisation, we were able to obtain the following results:

- Lasso Regression: Used LassoCV to find the optimal value of lambda as the convergence of GridSearchCV proved to yield unfruitful results due to the lack of convergence of the hyperparameters.
- Ridge Regression: Used GridSearchCV to loop through a space of ”alpha” values occupying from (0.001 to 2) and acquiring the best alpha for our regression.
- Decision Tree: Used GridSearchCV to loop through a space of ”splitter” values which included ”best”, ”random” while also looping through a space of ”max depth” values from (0

to 8).

### Modeling Event Details

Logistic regression or multi-layer perceptron classifiers (with an appropriate loss function, such as cross-entropy loss or Brier score) are able to produce a probability that the event following the most recent one will be of a particular type (e.g. a pass by the home team in a particular section of the field).

### Deducing Conditional Waiting Time Distributions

Given that the next event is of a specified type, it appears (see Figure 2) that the conditional waiting time distribution follows an exponential distribution.

In particular, the time from the most recent event until the next goal, conditional on it being scored by a particular team, follows an exponential distribution.

If we assume that any particular infinitesimal time period of $dt$ seconds has a particular constant probability $\lambda_H \cdot dt$ of containing a goal for the home team, a particular constant probability $\lambda_A \cdot dt$ of containing a goal for the away team, and a particular constant probability of not containing a goal, then we can model the unconditional waiting time $W$ until the next goal (regardless of which team it is scored by) as the minimum of the two conditional waiting time distributions ($W_H$ and $W_A$), which is again exponential with parameter $\lambda_H + \lambda_A$.

In this model, the probability that the next goal is scored by the home team is $P(W_H < W_A) = \frac{\lambda_H}{\lambda_H+\lambda_A}$, while the probability it is scored by the away team is $P(W_A < W_H) = \frac{\lambda_A}{\lambda_H+\lambda_A}$.

Observe that the unknown parameter $\lambda_A$ is equal to the quotient of expected unconditional waiting time $P(W_A < W_H)$ by $\mathbb{E}[W] = \frac{1}{\lambda_H+\lambda_A}$, which are the outputs of the event details model and waiting time model respectively, and so we can deduce that $\lambda_A = \frac{P(W_A<W_B)}{\mathbb{E}[W]}$ and similarly $\lambda_B = \frac{P(W_B<W_A)}{\mathbb{E}[W]}$.

Given appropriate loss functions, the models for probabilistic classification and waiting time estimation should respectively estimate probabilities and expectations as the amount of data used

grows (this property is sometimes referred to as Fisher-consistency, and if it is satisfied, the simulation output will also be Fisher-consistent). However, we do not have an explicit model for the error of this method of parameter estimation (not only is it a quotient of two random variables, but these random variables will be the outputs of two black-box machine learning models), nor for how this error will impact the quality of predictions. This is explored further in the discussion section.

The significance of having access to these parameters is discussed below.

**Synthesising An Outcome Distribution From Conditional Waiting Times**

We now have access to the conditional waiting time distributions $W_H|R \sim \text{Exp}(\lambda_H^{(R)})$ and $W_A|R \sim \text{Exp}(\lambda_A^{(R)})$, where $R$ represents information about the most recent event. Similarly, because of the Markov property, modeling the waiting time from the next goal to the goal after that is the same problem as modeling the waiting time from the beginning of the game to the first goal.

Considering the game outcome as a two-entry vector representing the scores for the home team and the away team, we may write the stochastic recurrence relation

$$G(t;\lambda_A^{(R)},\lambda_H^{(R)}) = P(A)\left(\begin{bmatrix}0\\1\end{bmatrix} + G(t-W_A;\lambda_A,\lambda_H)\right) + P(H)\left(\begin{bmatrix}1\\0\end{bmatrix} + G(t-W_H;\lambda_A,\lambda_H)\right),$$

where $G$ represents the number of goals scored after the most recent match event,

$$P(A) = P(W_A < t \wedge W_A < W_H|R) = P(W_A < W_H < t|R),$$

$$P(H) = P(W_H < t \wedge W_H < W_A|R) = P(W_H < W_A < t|R),$$

the variable $t$ represents the time left until the end of the match period, and $\lambda_A, \lambda_H$ are the values of each parameter at the beginning of the game (or equivalently, immediately after each goal), rather than the parameters derived from the most recent event.

The solution to this recurrence can then be approximated with the Monte Carlo method; the only random variables involved are $W_A$ and $W_H$, so we sample values for these at each step based on the known parameters, and then recursively evaluate the appropriate term in the recurrence relation until $t$ becomes nonpositive in order to simulate a possible pathway for the rest of the game. These simulated pathways then give an empirical probability distribution for the possible game outcomes.

While this explanation gives the main idea, there are two final practical considerations. Firstly, the team in control of the kick-off after a goal has been scored and the game state has been reset will be the team who the goal was scored against, and this information should be incorporated. Secondly, the game is typically split up into two 45-minute halves, rather than a single 90-minute period, with a state reset in between. It is relatively simple to amend the recurrence to account for these, but we leave this detail for the implementation of the algorithm.

**Recursive Probabilistic Simulation**

The above method ignores for simplicity the events that might happen between the most recent event and the next goal.

One might expect an explicit stochastic model of all events in the match to be more sophisticated when it comes to learning causal relationships between all the different kinds of match events, and intuitively it seems that such explicit models (i.e., simulations) ought to have lower sample complexity because they incorporate more prior information about the structure of the dynamical system being modeled.

Rather than having two possible event types $A$ and $H$ (as we did when modeling only goals), we now have a large number of different event types $T_i$, determined by categorical features such as `eventName`, `subEventName`, `tags`, etc., and associated parameters $\lambda_{T_i}$. As before, we can make use of a probabilistic classifier and a waiting time estimation model to deduce that

$$\lambda_{T_i} = \frac{P\left(T_i = \underset{T_j}{\operatorname{argmin}}\, W_{T_j}\right)}{\mathbb{E}[W]}.$$

To produce a probability distribution for the final score, we again make use of Monte Carlo simulation, but rather than modifying the score at each step, we modify it only when a goal occurs, and otherwise we simply modify the state of the game.

**Handling Numerical Features**

Because we have assumed that there are a finite number of exponential distributions from which we take the smallest realised outcome, our framework so far is unable to deal with continuous numerical features that serve to specify a particular state, because this would lead to an infinite number of possible 'next states'.

Inspired by the use of Kalman Filtering in control theory (see Figure 3 by Aimonen, 2011), we initially proposed solving this problem by first sampling an evolution of the categorical states, and then using the pair of old/new states along with the previous position to generate an estimate of the new expected position, alongside an estimate of the uncertainty about this position to be used for sampling a Gaussian noise term that is added to this expectation in simulation.

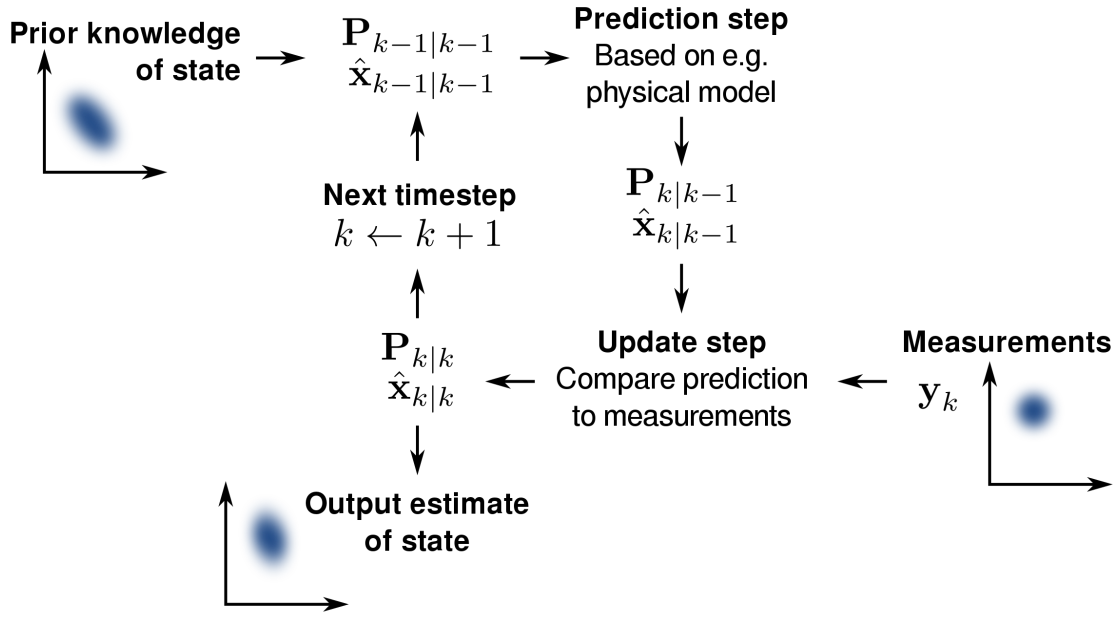


Figure 3. Diagram Explaining the Kalman Filtering Technique

Unfortunately, the appropriateness of a Kalman filter requires an assumption of multivariate nor-

mality. Since the playing field is bounded by a finite rectangle, the position of the next event certainly cannot follow a multivariate normal distribution. Even the change in position does not follow a normal distribution, as confirmed by a Shapiro-Wilk test applied to the marginals of the distribution ($\Delta x$ and $\Delta y$). See Figure 4 for visualisations of both of these.

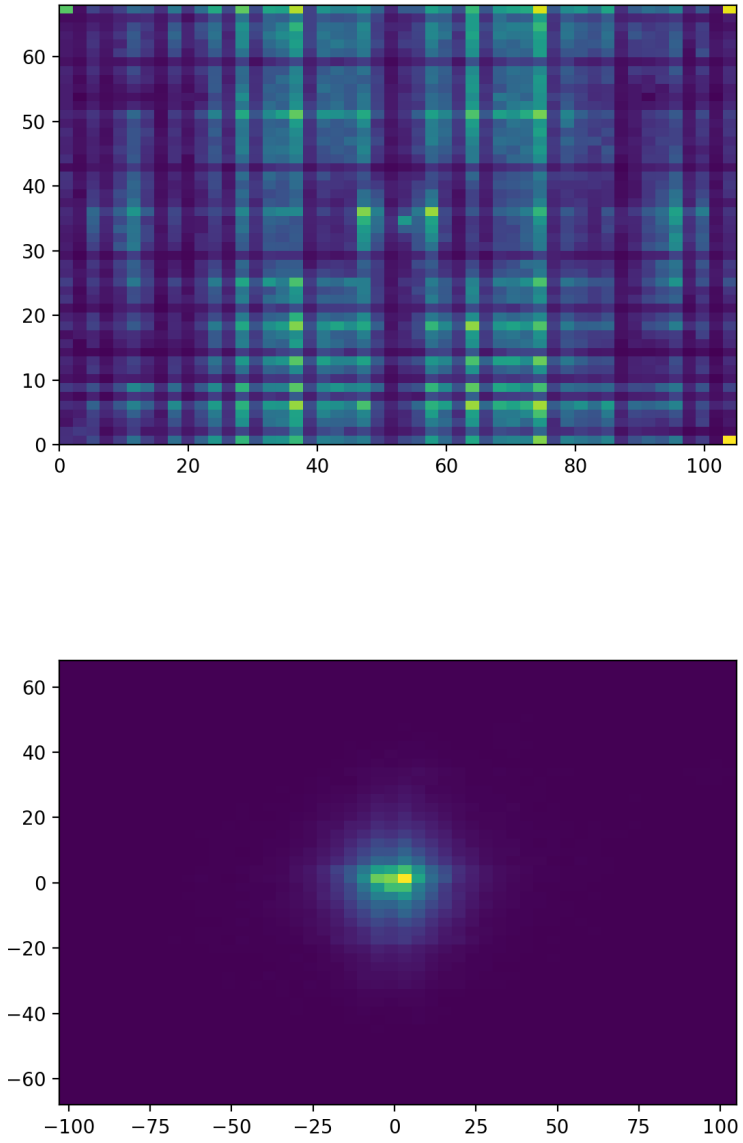


Figure 4. Heatmaps of Field Position and Field Position Change

If we had a method for sampling from a probability distribution given empirical moments (which exist and uniquely determine the distribution, since the distribution has compact support), we could train a model to estimate these moments conditional on the most recent event, and thereby modify the Kalman filter technique to handle non-normal data. Unfortunately, we are not aware of any such method, or of any appropriate alternative.

To remedy this, we instead divide the field into large sections and treat these as separate categorical variables in order to obtain a fully categorical representation of the state. While this reduces the granularity of position, and discards information about the spatial proximity of different sections of the field, it nonetheless solves the problem of making the state transition graph finite. In particular, the provided code splits the field three ways in each direction.

## EVALUATION

The most reasonable method for splitting the data into train and test sets seems to be training on matches prior to a particular date and testing on matches played after this date. Training on a random subset of the matches and testing on the others also seems reasonable, since there should not be strong causal interactions between different matches.

In principle, one could use pricing data from a real-world betting market to give a baseline for how the certainty of an efficient pricing model ought to update as the match progresses. Unfortunately, this is not data that is readily accessible, but if collected it would be a useful benchmark for model quality.

Although the component models of the more complex method represent very, very simple machine learning problems, due to miscommunications they were unfortunately not implemented, and so despite the rather nice high-level algorithm, we were unable to produce evaluable results for the high-level model (though results for the individual components were obtained and are discussed in the methodology section).

# DISCUSSION

## Limitations of The Method

The model is not able to easily incorporate information communicated by the actions of other bettors in a prediction market. It also depends on the various assumptions made in the methodology section. These assumptions being false would in itself be interesting, and potentially provide new insights into the nature of the underlying system (e.g. the non-independence of the conditional waiting time distributions). If the system does not satisfy the Markov property, improved feature engineering might serve to remedy this.

## Errors In Parameter Estimation

Although the simulation method described appears to be Fisher-consistent (as discussed in the methodology section), asymptotic consistency is not the only important metric to consider when evaluating an estimation procedure.

## Applications of The Stochastic Simulation Framework

The stochastic simulation framework described above is fairly general, and could be applied to other time series involving stochastic state-based dynamical systems with exponentially distributed waiting times. The key assumption is constancy of the rate parameter over time (independence and exponentiality follow as a corollary).

## Applications To Team Management

The state-based simulation model we have constructed might find applications in testing causal hypotheses informing the problems of team construction, player coaching, and understanding match events. In particular, the model might be extended to identify key drivers of match outcomes, as well as the consequences of players making different decisions or improving particular kinds of athletic skill or accuracy, as well as to identify the relative contribution of different players to match outcomes.

APPENDIX

| **eventName** | **subEventName** | **Frequency** |
|---|---|---|
| Duel | Air duel | 14448 |
| | Ground attacking duel | 20444 |
| | Ground defending duel | 20324 |
| | Ground loose ball duel | 11412 |
| Foul | Foul | 2930 |
| | Hand foul | 96 |
| | Late card foul | 25 |
| | Out of game foul | 34 |
| | Protest | 40 |
| | Simulation | 11 |
| | Time lost foul | 11 |
| | Violent Foul | 7 |
| Free Kick | Corner | 1546 |
| | Free Kick | 2808 |
| | Free kick cross | 644 |
| | Free kick shot | 127 |
| | Goal kick | 2399 |
| | Penalty | 28 |
| | Throw in | 6568 |
| Goalkeeper leaving line | Goalkeeper leaving line | 489 |
| Interruption | Ball out of the field | 10555 |
| | Whistle | 77 |
| Others on the ball | Acceleration | 1703 |
| | Clearance | 4510 |
| | Touch | 12892 |
| Pass | Cross | 4808 |
| | Hand pass | 980 |
| | Head pass | 7959 |
| | High pass | 9638 |
| | Launch | 3800 |
| | Simple pass | 97736 |
| | Smart pass | 2386 |
| Save attempt | Reflexes | 784 |
| | Save attempt | 464 |
| Shot | Shot | 3283 |

Table 1. Event frequencies in the England dataset

REFERENCES

Aimonen, P. (2011). Basic concept of kalman filtering. https://commons.wikimedia.org/wiki/File:Basic_concept_of_Kalman_filtering.svg

Lewis, M. (2010). *Moneyball: The art of winning an unfair game*. Norton Agency Titles.

Pelit, A. (2022). Mls signs $270m betting data deal with img arena. https://www.sportico.com/leagues/soccer/2022/mls-signs-betting-data-deal-img-arena-1234681696/

Taleb, N. N. (2020). Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv>stat*. https://arxiv.org/abs/2001.10488

Pappalardo et. al.. (2019). A public data set of spatio-temporal match events in soccer competitions. *Nature*. https://doi.org/10.1038/s41597-019-0247-7