Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186-5140
Volume 11, Number 1, pages 25–29, January 2022

Decision Making in American Football under State Uncertainty by Stochastic Inverse Reinforcement Leaning

Risa Takayanagi
The University of Electro-Communications
Tokyo, Japan

Keita Takahashi
The University of Electro-Communications
Tokyo, Japan

Masaya Wataba
The University of Electro-Communications
Tokyo, Japan

Kazunori Ohkawara
The University of Electro-Communications
Tokyo, Japan

Toma Sogabe*
The University of Electro-Communications
Tokyo, Japan
Grid, Inc.
Tokyo, Japan
sogabe@uec.ac.jp

***Abstract*—Reinforcement Learning (RL) techniques are often used to analyze and evaluate the strategies of virtual game to maximize a well-defined preset reward. However, in realistic sports game such as American football the reward functions are hardly definable. Meanwhile, how many yards are gainable on the next offence in real American football is also usually uncertain during strategy planning. In order to tackle these issues, we propose a stochastic inverse reinforcement leaning (IRL) algorithm. The expert data for IRL are built by using the American football 2017 season event data in National Football League (NFL). The stochastic state transition distribution is extracted from the same dataset. A mixture density network is used to learn the probabilistic distribution. At the last, simulation results from the maximum entropy IRL are compared with the ones from mathematical two-stage stochastic optimization.**

Keywords—Strategy evaluation, American football, Inverse reinforcement learning, Uncertainty, Mixture density network, Stochastic optimization,

I. Introduction

Reinforcement leaning (RL) technics have been applied in many fields to analyze and evaluate strategies and assistant decision making. For example, virtual games RL programs like AlphaGo[1], AlphaGo Zero[2] and Atari[3] have defeated humans players. Moreover, in sports fields like Ice Hockey[4], Soccer[5], and Basketball[6], RL technics were used to evaluate the decision-making and player behavior. However, realistic sports game such as in American football, the reward functions for value approximation are hardly definable[7]. Meanwhile, as opposed to maze problem, when agent act in each state it is uncertain for next state transition as the yard to be gained in American football. The uncertainties are varied by the condition of players, weather, and game place and etc. Thus the state transition becomes stochastic. Furthermore, under such a situation, it is difficult for the coach to select and decide tactics of each game in real play. The purpose of this work is aimed to tackle these problems by proposing a stochastic inverse reinforcement learning (IRL) algorithm. For our experiments, we choose maximum entropy (Max-Ent) IRL[8], in which the agents learn reward function from the expert demonstration data based on the professional play-by-play in NFL 2017 season. The dataset contains 9516 plays of 21 games by 32 teams. The information includes what kind of play choices each professional team have made under each situation. Our model learns from this data and learn a policy to decide which tactic is best to decide in uncertainly situations. We expand the Max-Ent IRL algorithm to handle the stochastic state transition by introducing Mixture Density Network (MDN) [9] to learn the probabilistic distribution of next state. At the last, simulation results from the maximum entropy IRL are compared with the ones from mathematical two-stage stochastic optimization.

II. Background

1. Max Entropy Inverse Reinforcement Learning

In Max Entropy IRL [8] reward function R represented by the frequency of the agent visits to states $s \in S$, where S is a finite set of states. The function is parameterized by some rewards weights θ.

$$R(s,\theta) = \sum_t \theta^T \cdot \emptyset(s_t) \qquad (1)$$

Here $\emptyset$ is the visiting trajectory. The reward function R is used to learn the policy π^*.

$$\pi^*(a|s_t) = argmax\left\{R(s_{t+1},\theta)+\gamma\sum_{s_{t+1}}\left(P(s_{t+1}|s_t,a)V^{\pi}(s_{t+1})\right)\right\} \quad (2)$$

The transitions probability of moving from state s to next state s' as a result of $\pi(a)$ is represented as $P(s'|s,a)$. For each states state s, V denote the state value under a policy π. The visiting frequency μ for each training T is defined as follows:

$$\mu_T(s_i,\theta)=\sum_{a\in A}\sum_{s\prime\in S}\mu_{T-1}(s_t)\pi^*(a|s_t)\,P(s_{t+1}|s_t,a) \quad (3)$$

The probability of visiting frequency for each state is given as a sum of visiting frequency μ.

$$\mathbb{p}(s_i|\theta)=\sum_{t=1}^{T}\mu_t \quad (4)$$

Expert visiting frequency μ' of each state is extracted from data. To update parameter θ is calculated using the difference of visiting frequency between expert reward and the result of $\mathbb{p}(s_i|\theta)$ based on policy π^*.

$$\theta \leftarrow \theta+\alpha\left(\mu'-\mathbb{p}(s_i|\theta)\right) \quad (5)$$

The updated θ is used to evaluate the weight of learning reward by formula (1).

2. Mixture Density Network

Mixture Density Network [9] is used to learn uncertain realistic models and has the merits of providing a probability distribution over a range of outputs given the input [10]. The technics learn multiple Gaussian distribution as follows. Given vector $\boldsymbol{x}$ of input, the probability of y given x $p(y|\boldsymbol{x})$ can be approximated as:

$$p(y|x)=\sum_{k=1}^{K}\pi_k(\boldsymbol{x})\,\mathcal{N}\left(y\middle|\mu_k(\boldsymbol{x}),\sigma_k^2(\boldsymbol{x})\right) \quad (6)$$

$$\sum_{k=1}^{K}\pi_k(\boldsymbol{x})=1 \quad (7)$$

Here k is the index of corresponding mixture component, they are up to K. The parameter $\pi_k(\boldsymbol{x})$ denote the mixing coefficients of k th Gaussian distribution. $\mathcal{N}$ means k th Gaussian distribution.

$$\mathcal{N}(y|\mu_k(\boldsymbol{x}),\sigma_k^2(\boldsymbol{x}))=\frac{1}{\sqrt{2\pi\sigma_k^2(\boldsymbol{x})}}\exp\left(-\frac{(\mu_k(\boldsymbol{x})-y)^2}{2\sigma_k^2(\boldsymbol{x})}\right)$$

Where $\mu_k(\boldsymbol{x})$ represents the mean of k th kernel, the k th variance described by $\sigma_k(\boldsymbol{x})$.

3. Mathematical two-stage Stochastic Optimization

Mathematical two-stage stochastic Optimization is generally used as a conventional algorithm to perform stochastic optimization [11]. The objective function is defined as:

$$Maxmize\sum_{n}p(s_n^{t+1}|a^t,s^t)V(s_n^{t+1})$$

where p is transition probability from s_t to s_{t+1} as a result of action a_t. $V(s)$ represent the value at the state $s\in S$. The state s defined as

$$s^i=\begin{bmatrix}c^i\\ r^i\end{bmatrix}=\begin{bmatrix}\text{floor}\left(\frac{i}{11}\right)\\ i\ mod\ 11\end{bmatrix}$$

III. American Football for IRL

1. The rules of American football

American football is one of the most popular sports in United States. As opposed to Soccer and Basketball in which the time flows continuously, American Football is independent for each play. Between each play, coach instructs 11 players in the field on how and what play would proceed for the next game. The decision making during this interval plays vital role and has direct effect on final scores.

Offense side has opportunity to play 4 times each sequence, they must advance 10 yards in these plays. Generally, the 4th play is used to recover their own position, thereby the offence team try to advance 10 yards in 3 plays as much as possible. If offence team advance 10 yards, they could get a new opportunity to play. In consequence, they aim to advance the ball to the goal line at the end of field. Figure 1 shows several examples of how the offence side advances. The first example (top) illustrating a success scenario to advance 10 yards in 3 plays. The middle on is failure scenario to advance 10 yards. As a result, offense team use the next play to recover their position and the opposite team will attack. The bottom is also a failure example to advance. In this case, defensive team steal the ball and that team will attack in the next play.

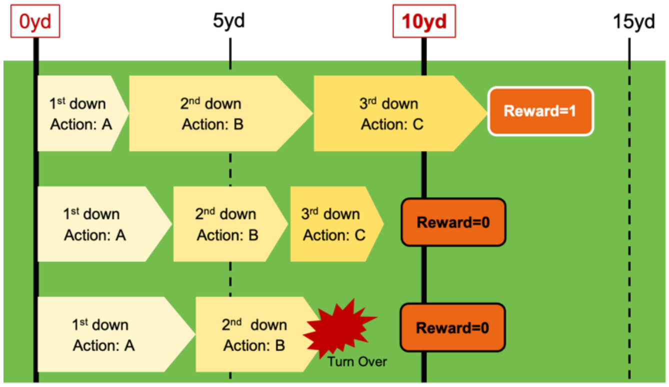


Figure 1: Examples of offense advance.

As a result, each play has different features and strategies for each team, which is correctly determined by the movement 11 players. However, plays can be categorized according to their features. Representing categories mainly includes pass play or run play. Pass play is the throw the ball from player to player, it has a low probability of success and high risk of turnover by the opponent team but has a chance to advance longer distance. In contrast, though run play is low risk of turnover and is difficult to advance long distance. The decision strategies involved in American football is primarily focused on how to well select and balance the two types of plays.

2. A Markov Model for American Football

In this section, we will describe how to map the American Football game into the stochastic IRL scheme. The Markov Decision Process (MDP) [12] is defined by a set of states S, and set of actions A, and transition functions $T: S \times A \rightarrow PD(S)$, where PD represents the set of probability distributions over finite set of state S. For an agent, there is the reward function $R: S \times A \rightarrow \mathbb{R}$, and have the discount factor $0 < \gamma < 1$ that discount the value of the state over time.

In this work we define the states based on the field of American football as shown in Table 1. Each state has 3 substates, therefore the number of state combination is $3^3 = 27$ states. 5 kinds of action are considered in this work: guard, tackle, end, short, deep, which is categorized according to the features of plays. Run plays are guard, tackle, and end, pass plays are short and deep.

Table1. States of MDP

State	Detail
Field position: Position of the field that play start. 100 yards line means the goal to get point.	Red: 0~20yds Yellow: 20~80yds Blue: 80~100yds
Down: How many times of out of 3	1st down 2nd down 3rd down
Distance: How long distance to advance 10 yds	Long: 7yd~ Middle: 4~6yd Short: ~3yd

3. Stochastic IRL for American Football

To tackle the problem of stochastic transition, we apply MDN to estimate the probability distribution of next state. Given the input of state and action, MDN learn the probability distribution of the gainable distance each situation. The probability values are given for each predicted gainable distance which we integer-discretized it from 0 to 10. In IRL, the value used as transition probability T to learn reward function. A pseudo-code of Stochastic Max-Ent IRL is given in Algorithm 1.

```
Algorithm 1 Stochastic Max-Ent IRL algorithm

Estimate transition probability $T$ from MDN
Initialize reward function parameter $\theta$
Initialize state function $F$(s)
Calcurate expert visiting frequency $\mu'$ from data
For episode = 1 to N do
    1.  Calculate reward function, obtain $R(s) = \theta \cdot F$(s)
    2.  Estimate maximum-likelihood policy $\pi^*$ from $R(s)$
    3.  Execute action $a \sim \pi^*$
        The next state is estimated from $T$ of MDN result
        Estimate visiting frequency $\mu$
    4.  Calculate the defference of visiting frequency
        $\Delta\mu = \mu' - \mu$
    5.  $\theta' = learning\ rate \cdot \Delta\mu$
    6.  Update $\theta \leftarrow \theta'$
end for
```

IV. Experiments

1. Stochastic IRL

At first, we estimate transition probability T using MDN from expert data. The input is states and action at the time t. MDN learns distribution of advanced distance under each combination of state and action. The learnt transition probability is embedded in IRL to learn visiting frequency. The result of MDN is used to estimate next state under the calculated policy from reward function. Due to the use MDN, the agent is able to explore the policy even under the stochastic situation.

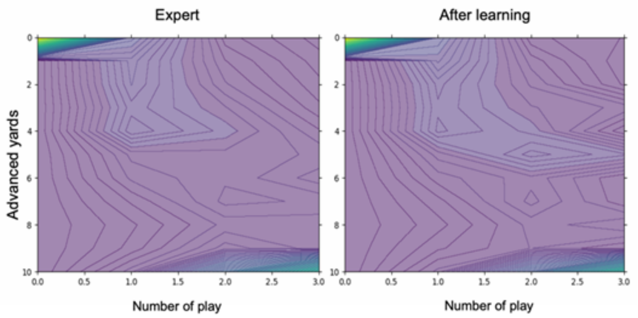


Figure 3. Contour lines of visit frequency in state space.

As a simple example, we set all expert action to “end”. The result of the difference in visiting frequency between expert and leaning agent shows Figure 3 as contour. Left side shows the visiting frequency from the expert demonstration while the right side shows the ones from the learnt agent. y axis represents the advanced yards accumulation. The figure shows that the agent is able to learn the reward function evidenced by a similar visiting trajectory mapping between the demonstrator and the learning agent. Upper left area in both figures represent the state of 1st down and 10 of distance that the start of the beginning of offense play in most games. Lower left area in both figures show the state after 3rd down, which means that visiting frequency is high because the goal is to advance 10 yards after 3rd down.

2. Mathematical Optimization for American Football

In Mathematical Optimization, we defined $V(s_{10}^{4}) = 1$, that represents the value that agent advanced 10 yards at the end of the 3rd play. We calculated tha value in all states and estimated the optimal action. In executing the calculation, we apply the transition probability from NFL event data as shown in Figure 3. The x label of yard gained represents the probability of advance to each yard from 0 yard to 10 yards after executing each action regardless of states. The probability is calculated from NFL data. If the really gained yard is minus, it is regarded as 0, and if the really gained yard is 10 or more, it is regarded as 10. This is attribute to the goal that is advanced by setting 10 yards or more.

The result of two-stage stochastic optimization is shown Figure 4. t means the number of plays of the 3 times. State means the remained distance to 10 yards. For example, in 1st

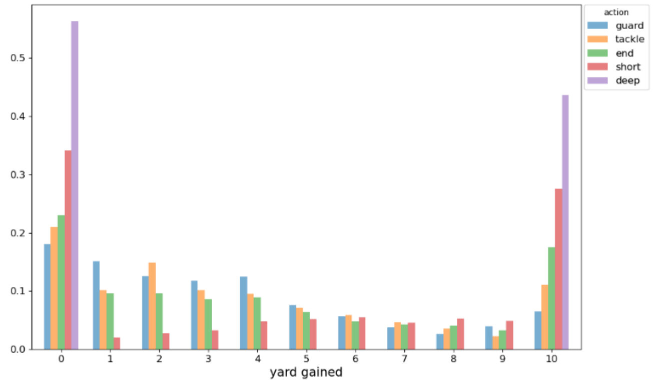
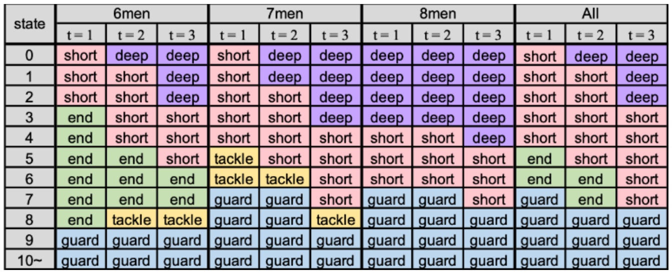


Figure 3: The advance probability of each action.

down and 10 yards of distance, we index the the matrix element as (state=0,t=1). Action has 5 types (e.g., guard, tackle, end, short, deep), and in each state, the best action calculated is highlighted. Top row (e.g., 6men, 7men, 8men, All) means the number of defense player in the tackle box. The more player in the tackle box, the better to defeat run play (e.g., guard, tackle, end). "All" is that considers all defensive pattern. For instance, in "6men" defense, 1st down, and 10 of distance, we confirm the matrix of 6men, state=0,t=1, which suggests that it is better to choose "short" play. We calculated for typical defense tactics and all tactics, that represent top row (e.g., 6men, 7men). The column of "All" is the result of optimizing regardless of defense tactics. As a whole, in short distance, the results suggests that it is better to choose run play (e.g., guard, tackle, end). The results are consistent with the perspective of domain knowledge. In case of evaluating by defense strategy, the smaller the number of players, the more choices the pass play. Generally, because the number of players is proportional to the tightness of defense for pass play, and if the number of players in tackle box is small, it is better to choose run play. Thus, the result is close to domain knowledge. However, there is a problem of flexibility, because the policy are greedy in the sense that only one best action is presented from the mathematical optimization. In real game, even under the same state, which play the coach decide shows a lot of uncertainties and are heavily dependent on ad-hoc condition and situtaion during each play such as scores, player condition and previous tactics.



state	6men			7men			8men			All		
	t=1	t=2	t=3	t=1	t=2	t=3	t=1	t=2	t=3	t=1	t=2	t=3
0	short	deep	deep	short	deep	deep	deep	deep	deep	short	deep	deep
1	short	short	deep	short	deep	deep	deep	deep	deep	short	short	deep
2	short	short	deep	short	short	deep	deep	deep	deep	short	short	deep
3	end	short	short	short	short	deep	deep	deep	deep	short	short	short
4	end	short	short	short	short	short	short	short	deep	short	short	short
5	end	end	short	tackle	short	short	short	short	short	end	short	short
6	end	end	end	tackle	tackle	short	short	short	short	end	end	short
7	end	end	end	guard	guard	short	guard	guard	short	guard	end	short
8	end	tackle	tackle	guard	guard	tackle	guard	guard	guard	guard	guard	guard
9	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard
10~	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard	guard

Figure 4: The result of optimal action each state based on mathematical optimization.

V. Conclusion

In this work, we proposed a method to use IRL for American football as an example of under the stochastic uncertainly situation. The results show that the agent is able to learn the reward function evidenced by a similar trajectory mapping between the demonstrator and the learning agent. The mathematical optimization shows also ability to generate reasonable strategies under specified condition. However, a reliable experiment using MDN are not yet for both approaches thus generation ability to deal with realistic uncertain is poor for both approaches and is an undergoing subject. Meanwhile, a Bayesian IRL instead of Max-Ent IRL is also under investigation for this purpose. We will describe the latest progress and result regarding these issues at workshop.

VI. Reference

[1] D Silver, A Huang, C Maddison, A Guez, L Sifre, G Driessche, J Schrittwieser, I Antonoglou, V Panneershelvam, M Lanctot, S Dieleman, D Grewe, J Nham, N Kalchbrenner, I Sutskever, T Lillicrap, M Leach, K Kavukcuoglu, T Grapel, D Hassabis, "Mastering the game of Go with deep neural networks and tree search", Nature, 2016

[2] D Silver, J Schrittwieser, K Simonyan, I Antonoglou, A Huang, A Guez, T Hubert, L Baker, M Lai, A Bolton, Y Chen, T Lillicrap, Fan Hui, L Sifre, G van den Driessche, T Graepel, D Hassabis, "Mastering the game of Go without human knowledge", Nature, 2016

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller, "Playing Atari with Deep Reinforcement Learning", arXiv:1312.5602, 2013

[4] Guiliang Liu, Oliver Schulte, "Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation", *In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 3442–3448. International Joint Conferences on Artificial Intelligence Organization*, 2018.

[5] Liu, G., Luo, Y., Schulte, O., Kharrat, T. "Deep soccer analytics: learning an action-value function for evaluating soccer players", *Data Mining and Knowledge Discovery*, 2020

[6] Sandholtz, N. and Bornn, L. Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball. *The annals of applied statistics, 14(3):1122–1145,* 2020.

[7] Yudong Luo, Oliver Schulte, Pascal Poupart, "Inverse Reinforcement Learning for Team Sports: Valuing Actions and Players", *In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence IJCAI-20, pp. 3356-3363. International Joint Conferences on Artificial Intelligence Organization*, 2020.

[8] Brian D. Ziebart, Andrew Maas, J.Andrew Bagnell, and Anind K. Dey, "Maximum Entropy Inverse Reinforcement

Learning", *in AAAI Conference on Artificial Intelligence*, 2008.

[9] Christopher M. Bishop, "Mixture density networks", Technical report, Aston University, 1994.

[10] Oliver Borchers, "A Hitchhiker's Guide to Mixture Density Networks", towards data science, 26 Oct 2021.URL https://towardsdatascience.com/a-hitchhikers-guide-to-mixture-density-networks-76b435826cca

[11] Takayuki Shiina, "Stochastic Programming Model for Unit Commitment Problem", *RIMS Kokyuroku*, 2003.

[12] Howard, Ronald A. "Dynamic Programming and Markov Processes", *The MIT Press, Cambridge, Massachusetts,* 1960