

Learning to Collaborate: Multi-Scenario Ranking via Multi-Agent Reinforcement Learning*

Jun Feng^{1,†}, Heng Li^{2,†}, Minlie Huang^{1,*}, Shichen Liu^{2,*}, Wenwu Ou²,
Zhirong Wang², Xiaoyan Zhu¹

¹State Key Lab on Intelligent Technology and Systems, Tsinghua National Lab for Information Science and Technology

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

²Alibaba Group, Hangzhou, China

feng-j13@mails.tsinghua.edu.cn;heng.li@alibaba-inc.com;aihuang@tsinghua.edu.cn;shichen.lsc@alibaba-inc.com;
santong.oww@taobao.com;qingfeng@alibaba-inc.com;zxy-dcs@tsinghua.edu.cn

ABSTRACT

Ranking is a fundamental and widely studied problem in scenarios such as search, advertising, and recommendation. However, joint optimization for multi-scenario ranking, which aims to improve the overall performance of several ranking strategies in different scenarios, is rather untouched. Separately optimizing each individual strategy has two limitations. The first one is **lack of collaboration between scenarios** meaning that each strategy maximizes its own objective but ignores the goals of other strategies, leading to a sub-optimal overall performance. The second limitation is the **inability of modeling the correlation between scenarios** meaning that independent optimization in one scenario only uses its own user data but ignores the context in other scenarios.

In this paper, we formulate multi-scenario ranking as a fully cooperative, partially observable, multi-agent sequential decision problem. We propose a novel model named Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) which has a communication component for passing messages, several private actors (agents) for making actions for ranking, and a centralized critic for evaluating the overall performance of the co-working actors. Each scenario is treated as an agent (actor). Agents collaborate with each other by sharing a global action-value function (the critic) and passing messages that encodes historical information across scenarios. The model is evaluated with online settings on a large E-commerce platform. Results show that the proposed model exhibits significant improvements against baselines in terms of the overall performance.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; • **Theory of computation** → **Multi-agent reinforcement learning**;

*Corresponding authors: Minlie Huang, aihuang@tsinghua.edu.cn; Shichen Liu, shichen.lsc@alibaba-inc.com

†The authors contributed equally to this study.

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW 2018, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04.

<https://doi.org/10.1145/3178876.3186165>

KEYWORDS

multi-agent learning, reinforcement learning, learning to rank, joint optimization

ACM Reference Format:

Jun Feng^{1,†}, Heng Li^{2,†}, Minlie Huang^{1,*}, Shichen Liu^{2,*}, Wenwu Ou², Zhirong Wang², Xiaoyan Zhu¹. 2018. Learning to Collaborate: Multi-Scenario Ranking via Multi-Agent Reinforcement Learning. In *Proceedings of The 2018 Web Conference (WWW 2018)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3178876.3186165>

1 INTRODUCTION

Nowadays, most large-scale online platforms or mobile Apps have multiple scenarios that may involve services such as search, advertising, and recommendation. There are some well-known platforms of different kinds. Taobao is an E-commerce platform where users can search for and buy products through querying, bookmarking, or recommendation. Yahoo! is a comprehensive web site where users can read news, watch movies, make shopping, and more. One of the common features of these services is that ranking strategy serves as a fundamental function to provide a list of ranked items to users. Machine learning techniques have been widely applied in optimizing these ranking strategies [8, 28, 32, 50] to facilitate better services for search, advertising, or recommendation.

However, ranking strategy in one scenario only optimizes its own metric, without considering the correlation between scenarios (or applications). **In these platforms, strategies in different scenarios may be developed by different teams, and optimized by different methods with different metrics.** Such metrics may include Click Through Rate (CTR), Conversion Rate (CVR), and Gross Merchandise Volume (GMV). However, separate optimization of single scenario cannot guarantee the globally optimal performance of the entire platform. Instead, if the strategies in different scenarios can work collaboratively, we can expect a better overall performance. Let's illustrate this with a toy example. In a long beach, as shown in Figure 1, there are two sellers (denoted by A and B), located at different positions for selling their snacks. The top figure indicates the initial location, where people on the left side of the beach buy snacks at A and people on the right at B. The middle figure shows that when A moves right, he can sell more snacks (A can cover more people than B). Similar cases to B. The bottom figure indicates an optimal solution to this non-cooperative game, where the two sellers compete with each other and they are both at the center

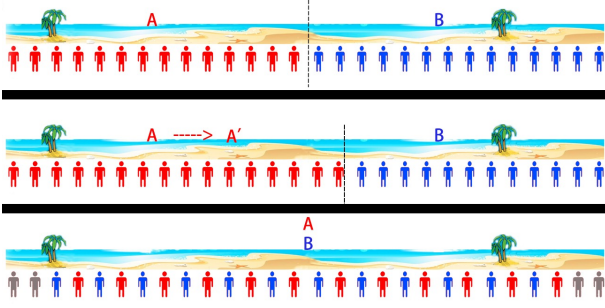


Figure 1: A competitive game for two sellers (A and B) selling snacks in a long beach. The top figure shows the initial location, the middle one shows the competing process, and the bottom one shows a solution when the two sellers are competitors. People in red are likely to buy snacks at A, and people in blue at B. People in grey are those beyond the scope of A and B.

of the beach. However, this is a definitely sub-optimal solution if we want to optimize the total income of the two sellers, as some people (in grey) are beyond the scope of them.

This simple example demonstrates that collaboration between scenarios in a system is extremely important if the objective is to optimize the total return of the system. This is also the case for E-commerce platforms which have many different scenarios in service. In a large E-commerce platform, we indeed observed competitor behaviors: increasing CTR in product search drops that in search advertisement systems, and increasing GMV in main search (the entrance search service of the system) may drop that in-shop search (the search service within a certain shop). The famous Cournot model [11] can be another example, denoting that if there are more than one oligarch in the market, the total revenue becomes more if the oligarchs are cooperative with each other, but less if they are competitive. When ranking strategies in different scenarios are optimized independently but not collaboratively, each strategy maximizes its own objective but ignores the goals of other strategies, leading to a sub-optimal overall performance. We term this issue as the **lack of collaboration between scenarios**.

Another limitation caused by independent optimization exists in the **inability of modeling the correlation between scenarios**. The user behaviors in different scenarios are correlated and indicative of what they are looking for, which is valuable for optimizing ranking algorithms. Our investigation, on a corpus which consists of user logs of millions of users from Taobao (a large E-commerce platform in China), shows that 25.46% switches from main search to in-shop search and 9.12% switches from in-shop search to main search. In addition, scenario switch not only happens between main search and in-shop search, but also among other scenarios such as main search, advertising, and recommendation. Undoubtedly, independent optimization in one scenario only uses the partial information (the data within its own scenario) of the user behavior data, which may lead to suboptimal performance.

In order to deal with the above limitations, we propose a novel model for joint multi-scenario ranking in this paper. The model

jointly optimizes ranking strategies for different scenarios through collaboration. In detail, different ranking strategies in a system share an identical goal. The ranking results in one scenario are based on the previous ranking results and user behaviours from all other scenarios. In this way, the ranking strategies collaborate with each other by sharing the same goal; and since each strategy has access to all historical user data across different scenarios, the algorithm within a scenario can make full use of the complete user context.

We cast the multi-scenario ranking task as a fully cooperative, partially observable, multi-agent sequential decision problem. The sequential process works as follows: a user enters a scenario, and browses, clicks or buys some items, and then the search system (the model) changes its ranking strategy by adjusting the ranking algorithm when the user navigates into a new scenario or issues a new request. The process is repeated until the user leaves the system. Thus, the current ranking decision definitely affects the following decisions.

We propose a novel model named Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG). Each ranking strategy in one scenario is treated as an agent. Each agent takes local observations (user behavior data) and makes local actions for ranking items with its private actor network. Different agents share a global critic network to enable them to accomplish the same goal collaboratively. The critic network evaluates the future overall rewards starting from a current state and taking actions. The agents communicate with each other by sending messages. The messages encode historical observations and actions by a recurrent neural network such that agents have access to all historical information. In this manner, our model can optimize ranking strategies in multiple scenarios jointly and collaboratively, and utilizes the complete user behavior data across different scenarios.

The contributions of this paper include:

- We formulate multi-scenario ranking (or optimization) as a fully cooperative, partially observable, multi-agent sequential decision problem.
- We propose a novel, general multi-agent reinforcement learning model named Multi-Agent Recurrent Deterministic Policy Gradient. The model enables multiple agents (each corresponding to a scenario) to work collaboratively to optimize the overall performance.
- We evaluate the model with online settings in Taobao, a large online E-commerce platform in China. Results show our model has advantages over strong baselines (Learning-to-rank models).

2 BACKGROUND

2.1 Ranking Strategy

Learning to rank (L2R) [33] has been widely applied to deploy ranking strategies in many online platforms. The basic idea of L2R models is that the ranking strategy can be learned and optimized using a set of training samples. Each sample consists of a query and a ranked list of items/documents relevant to that query. The ranking function computes a score for each item with a set of features. The parameters of the ranking function can be learned by

various algorithms, such as point-wise [15, 29], pair-wise [2, 37], and list-wise methods [3, 6].

2.2 Reinforcement Learning

Reinforcement learning[42] is a framework that enables an agent to learn through interactions with the environment. At each step t , an agent receives the observation o_t of the environment, and takes an action a_t based on a policy μ . The environment changes its state s_t , and sends a reward r_t to the agent. The goal of the agent is to find a policy that maximizes the expected cumulative discounted reward $R(s_t, a_t) = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \dots$, where γ is a discount factor. Generally speaking, reinforcement learning methods can be classified into several branches, including policy-based [43], value-based [35], and actor-critic [25] which combines the two.

Next, we will give a brief introduction to DDPG [30] and DRQN [18] models, which are closely related to our proposed model.

2.2.1 DDPG. Deep Deterministic Policy Gradient (DDPG) is an actor-critic approach, which can be applied to solve the continuous action problems. DDPG maintains a policy function $\mu(s_t)$ and an action-value function $Q(s_t, a_t)$, which are approximated by two deep neural networks respectively, actor network and critic network. The actor network $\mu(s_t)$ deterministically maps a state to a specific action: $a_t = \mu(s_t)$. The critic network $Q(s_t, a_t)$ estimates the future cumulative rewards after taking action a_t at state s_t . In this paper, we employ a deterministic policy where the actor network outputs $a_t = \mu(s_t)$ which corresponds to the weight of a particular feature in a ranking algorithm. In other words, the action in our model is continuous, and DDPG is thus applicable.

2.2.2 DRQN. In real-world applications, the state of the environment may be partially observed. The agent is unable to observe the full state of the environment. Such a setting is called partially observable. Deep Recurrent Q-Networks (DRQN) are introduced to address the partial observation problem by considering the previous context with a recurrent structure. DRQN uses a Recurrent Neutral Network architecture to encode previous observations before the current timestep. Instead of estimating the state-action value function $Q(s_t, a_t)$ in Deep Q-Networks [35], DRQN estimates $Q(h_{t-1}, o_t, a_t)$, where h_{t-1} is the hidden state of the RNN which encodes the information of previous observations o_1, o_2, \dots, o_{t-1} . The recurrent network essentially applies this function to update its hidden states: $h_t = g(h_{t-1}, o_t)$ where g is a non-linear function.

2.3 Multi-Agent Reinforcement Learning

In multi-agent reinforcement learning (MARL) problems [4, 22, 31, 36], there are a group of autonomous, interacting agents sharing a common environment. Each agent receives their individual observations and rewards when taking an action based on each individual policy function. The agents can be fully cooperative, fully competitive, or with mixed strategies. Fully cooperative agents share a common goal and maximize the same expected return. Fully competitive agents have private goals opposite to each other (for instance, zero-sum games). Mixed strategies are in between the two extremes.

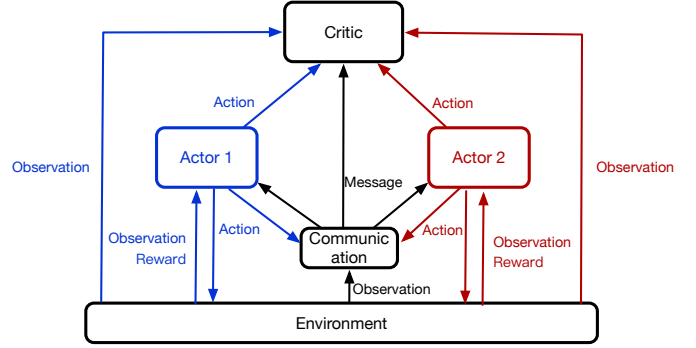


Figure 2: Overall model architecture. The model has a centralized, global critic network to evaluate the overall rewards. A communication module is used to generate messages that are shared among actors. Messages encode historical observations and actions, and can be used to approximate the global state of the environment. Each actor network represents an agent which receives its own local observations and a communication message, and makes private actions.

3 METHOD

To alleviate the two issues mentioned in the introduction section, we jointly optimize the ranking algorithms in multiple scenarios to maximize the overall returns by casting the task as a multi-agent reinforcement learning problem. We propose a novel model, named Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG). In this model, a ranking strategy in one scenario corresponds to an agent, and agents collaborate with each other to accomplish the same goal that optimizes the overall performance.

3.1 Problem Description

We formulate this task as a fully cooperative, partially observable, multi-agent sequential decision problem. More specifically:

Multi-Agent: there exist multiple ranking strategies/algorithms for different scenarios in a system. Each agent represents a ranking strategy and learns its own policy function which maps a state to a specific action.

Sequential Decision: users sequentially interact with the system. Thus, the agent actions are also sequential. At each step, the agent, which represents the scenario interacting currently with the users, chooses an action to respond to the user through a sorted list of items. The current actions affect the following actions in the future.

Fully Cooperative: all agents are fully cooperative to maximize a shared metric. Moreover, the agents pass messages to each other for communication, and the overall performance of these agents are evaluated by a centralized critic.

Partially Observable: The environment is partially observable, and each agent only receives a local observation instead of observing the full state of the environment.

3.2 Model

We design a Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) model to address the fully cooperative, partially observable, multi-agent sequential decision problem.

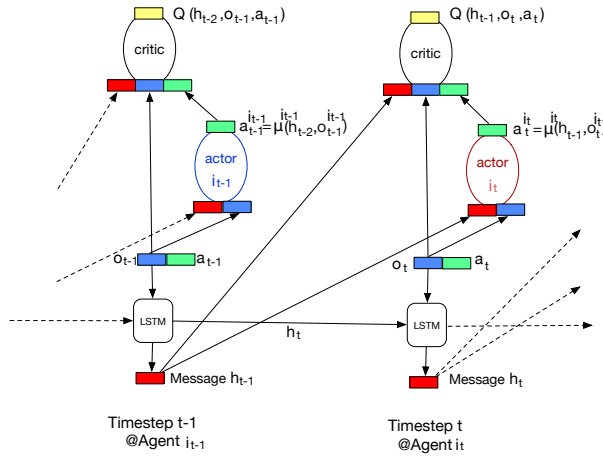


Figure 3: Detailed structure of MA-RDPG. The centralized critic network estimates the action-value function $Q(h_{t-1}, o_t, a_t)$ which indicates the future overall rewards when taking action a_t upon observing message h_{t-1} and observation o_t . The actor network outputs a deterministic action with $a_t^i = \mu^i(h_{t-1}, o_t^i)$ given the message and local observation as input. The messages are updated by a communication component which takes as input the observation o_t and action a_t . Red: Message; Blue: Observation; Green: Action.

3.2.1 Overview. Figure 2 shows the overall architecture of our model. For simplicity, we consider the case with two agents, each agent representing a scenario or strategy to be optimized. Inspired by DDPG [30], our model is built on top of the actor-critic approach [25]. We design three key modules to enable the agents to collaborate with each other: a centralized critic, private actors, and a communication component. The centralized critic evaluates an action-value function that indicates the expected future rewards for all agents taking actions from the current state. Each agent is represented by an actor network which maps a state to a specific action with a deterministic policy. Actions made by each actor network will be used for the agent to perform optimization in its own scenario.

We design a communication component using a Long Short-Term Memory (LSTM) architecture [21]. The LSTM encodes all local observations and the actions of all agents into a message vector. The message will be sent between agents for collaboration. Thanks to this component, the decision of each agent depends not only on its own previous observations and actions, but also on other agents' observations and actions. In addition, the messages can help the agents approximate the full state of the environment, which enables them to act more efficiently.

3.2.2 Model Details. A general reinforcement learning problem has a sequence of experiences $(o_1, r_1, a_1, \dots, a_{t-1}, o_t, r_t)$ where $o/r/a$ correspond to observation/reward/action respectively. As aforementioned, the environment in our problem is partially observable. In other words, the state s_t is the summary of the previous

experiences: $s_t = f(o_1, r_1, a_1, \dots, a_{t-1}, o_t, r_t)$ ¹. We are considering the problem with N agents $\{A^1, A^2, \dots, A^N\}$, each agent corresponding to a particular optimization scenario (ranking, recommendation, etc.). In this multi-agent setting, the state of the environment (s_t) is global, shared by all agents, while the observation ($o_t = (o_t^1, o_t^2, \dots, o_t^N)$), the action ($a_t = (a_t^1, a_t^2, \dots, a_t^N)$), and the intermediate reward ($r_t = (r_t(s_t, a_t^1), r_t(s_t, a_t^2), \dots, r_t(s_t, a_t^N))$) are all private, only possessed by each agent itself.

More specifically, each agent A^i takes action a_t^i with its own policy specified by $\mu^i(s_t)$, and obtains a reward $r_t^i = r(s_t, a_t^i)$ from the environment which changes its current state s_t to the next state s_{t+1} . In our task, all agents are collaborating to achieve the same goal. This leads to a collaborative setting of multi-agent reinforcement learning. We have a centralized action-value function $Q(s_t, a_t^1, a_t^2, \dots, a_t^N)$ (as critic) to evaluate the future overall return when taking the actions $(a_t^1, a_t^2, \dots, a_t^N)$ at the current state. We also have a global state representation of the environment, and each agent is represented by a private actor which observes local observations and takes private actions. Thus, the model belongs to an actor-critic reinforcement learning approach with a centralized critic and several private actors (each actor plays its role as an agent).

As shown in Figure 3, at step t , agent A^{i_t} receives a current local observation $o_t^{i_t}$ from the environment. The global state of the environment, shared by all agents, depends not only on all the historical states and actions of all agents in the sequential decision process, but also the current observation o_t . In other words, $s_t = f(o_1, a_1, \dots, a_{t-1}, o_t)$ ². To this end, we design a communication component using LSTM to encode the previous observations and actions of all agents into a message vector. With the message h_{t-1} sent between agents, the full state can be approximated as $s_t \approx \{h_{t-1}, o_t\}$ since the message h_{t-1} has encoded all previous observations and actions (see soon later). Agent A^{i_t} chooses the action $a_t^{i_t} = \mu^{i_t}(s_t) \approx \mu^{i_t}(h_{t-1}, o_t^{i_t})$ with the purpose of maximizing the future overall rewards estimated by the centralized critic $Q(s_t, a_t^1, a_t^2, \dots, a_t^N)$. Note that at each timestep, $o_t = (o_t^1, o_t^2, \dots, o_t^N)$ consisting of observations by all agents.

Communication Component We design a communication component to make the agents collaborate better with each other by sending messages. The message encodes the local observation and the actions at previous steps. At step t , agent A^{i_t} receives a local observation $o_t^{i_t}$ and a message h_{t-1} from the environment. The communication component generates a new message h_t taking as input the previous message h_{t-1} and current observation o_t . An agent can share the information with other collaborators through the message. As shown in Figure 4, we apply a LSTM architecture for this purpose. Formally, the communication component works as follows:

$$h_{t-1} = \text{LSTM}(h_{t-2}, [o_{t-1}; a_{t-1}]; \psi) \quad (1)$$

Note that o_t and a_t consists of observations and actions of all agents respectively, and each action a_t^i is also a real-valued vector since our problem is a continuous action reinforcement learning problem.

With the help of the message h_{t-1} , agents have access to an approximate of the full state of the environment: $s_t \approx \{h_{t-1}, o_t\}$,

¹In a fully observable environment, $s_t = f(o_t)$.

²Intermediate rewards r_t can be omitted in general for state representation.

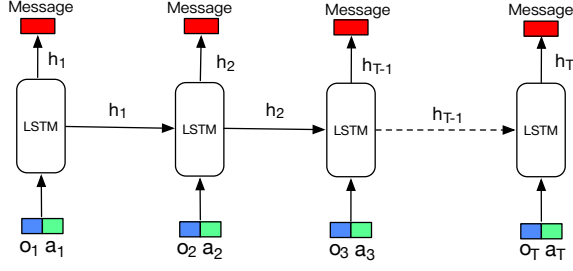


Figure 4: Communication component. The previous observations (o_t) and actions (a_t) are all taken as input to the LSTM network. The hidden states (h_{t-1}) are treated as messages which will be sent between agents. Note that o_t, a_t are vectors.

as an agent only receives its current observation o_t^i but not the full state s_t of the environment.

Private Actor Each agent has a private actor which receives local observations and shared messages, and makes its own actions. Since we deal with continuous action problems, we define the agent’s action as a vector of real values, $a^i = (w_1^i, \dots, w_{N^i}^i)$, $a^i \in \mathbb{R}^{N^i}$. Therefore, an action is a N^i -dimension vector, and each dimension is a continuous value. The action vector will be used in ranking algorithms or to control robots.

Since this is a continuous action problem which can be commonly seen in control problems [20, 30, 38], we resort to using a deterministic policy instead of a stochastic policy. The actor of each agent $\mu^i(s_t; \theta^i)$, parameterized by θ^i , specifies a deterministic policy that maps states to a specific action. At timestep t , agent A^{i_t} takes an action with its own actor network:

$$a_{t_t}^{i_t} = \mu^{i_t}(s_t; \theta^{i_t}) \approx \mu^{i_t}(h_{t-1}, o_{t_t}^{i_t}; \theta^{i_t}) \quad (2)$$

where $s_t \approx \{h_{t-1}, o_t\}$ as discussed in the communication component. In this manner, the actor is conditioned on the message h_{t-1} and its own current local observation $o_{t_t}^{i_t}$.

Centralized Critic Following DDPG, we design a critic network estimating the action-value function to approximate the expected future total rewards. As all agents share the same goal, we use a centralized critic $Q(s_t, a_t^1, a_t^2, \dots, a_t^N; \phi)$ to estimate the future overall rewards obtained by all agents after taking action $a_t = \{a_t^1, \dots, a_t^N\}$ at state $s_t \approx \{h_{t-1}, o_t\}$.

The above formulation is general and applicable to many agents that are alive all the time. In our setting³, there is only one agent A^{i_t} activated at timestep t , and $o_t = \{o_t^{i_t}\}$ and $a_t = \{a_t^{i_t}\}$. Hereafter, we will simplify the action-value function as $Q(h_{t-1}, o_t, a_t; \phi)$ and policy function as $\mu^{i_t}(h_{t-1}, o_t; \theta^{i_t})$.

3.3 Training

The centralized critic $Q(h_{t-1}, o_t, a_t; \phi)$ is trained using the Bellman equation as in Q-learning [48]. We minimize the below loss:

$$L(\phi) = \mathbb{E}_{h_{t-1}, o_t} [(Q(h_{t-1}, o_t, a_t; \phi) - y_t)^2] \quad (3)$$

³Because a user can be in only one physical scenario at each timestep.

ALGORITHM 1: MA-RDPG

```

Initialize the parameters  $\theta = \{\theta^1, \dots, \theta^N\}$  for the  $N$  actor
networks and  $\phi$  for the centralized critic network.
Initialize the replay buffer  $R$ 
for each training step  $e$  do
  for  $i = 1$  to  $M$  do
     $h_0 = \text{initial message}, t = 1$ 
    while  $t < T$  and  $o_t \neq \text{terminal}$  do
      Select the action  $a_t = \mu^{i_t}(h_{t-1}, o_t)$  for the active
      agent  $i_t$ 
      Receive reward  $r_t$  and the new observation  $o_{t+1}$ 
      Generate the message  $h_t = \text{LSTM}(h_{t-1}, [o_t; a_t])$ 
       $t = t + 1$ 
    end
    Store episode  $\{h_0, o_1, a_1, r_1, h_1, o_2, r_2, h_3, o_3, \dots\}$  in  $R$ 
  end
  Sample a random minibatch of episodes  $B$  from replay
  buffer  $R$ 
  foreach episode in  $B$  do
    for  $t = T$  downto  $1$  do
      Update the critic by minimizing the loss:
       $L(\phi) = (Q(h_{t-1}, o_t, a_t; \phi) - y_t)^2$ , where
       $y_t = r_t + \gamma Q(h_t, o_{t+1}, \mu^{i_{t+1}}(h_t, o_{t+1}); \phi)$ 
      Update the  $i_t$ -th actor by maximizing the critic:
       $J(\theta^{i_t}) = Q(h_{t-1}, o_t, a; \phi)|_{a=\mu^{i_t}(h_{t-1}, o_t; \theta^{i_t})}$ 
      Update the communication component.
    end
  end
end

```

where

$$y_t = r_t + \gamma Q(h_t, o_{t+1}, \mu^{i_{t+1}}(h_t, o_{t+1}); \phi) \quad (4)$$

The private actor is updated by maximizing the expected total rewards with respect to the actor’s parameters. If agent A^{i_t} is active at step t , the objective function is:

$$J(\theta^{i_t}) = \mathbb{E}_{h_{t-1}, o_t} [Q(h_{t-1}, o_t, a; \phi)|_{a=\mu^{i_t}(h_{t-1}, o_t; \theta^{i_t})}] \quad (5)$$

Following the chain rule, the gradients of the actor’s parameters are given as below:

$$\begin{aligned}
& \nabla_{\theta^{i_t}} J(\theta^{i_t}) \\
& \approx \mathbb{E}_{h_{t-1}, o_t} [\nabla_{\theta^{i_t}} Q(h_{t-1}, o_t, a; \phi)|_{a=\mu^{i_t}(h_{t-1}, o_t; \theta^{i_t})}] \\
& = \mathbb{E}_{h_{t-1}, o_t} [\nabla_a Q(h_{t-1}, o_t, a; \phi)|_{a=\mu^{i_t}(h_{t-1}, o_t)} \nabla_{\theta^{i_t}} \mu^{i_t}(h_{t-1}, o_t; \theta^{i_t})] \quad (6)
\end{aligned}$$

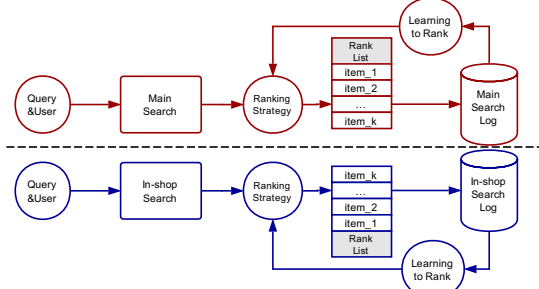
The communication component is trained by minimizing:

$$\begin{aligned}
& L(\psi) \\
& = \mathbb{E}_{h_{t-1}, o_t} [(Q(h_{t-1}, o_t, a_t; \phi) - y_t)^2 |_{h_{t-1}=\text{LSTM}(h_{t-2}, [o_{t-1}; a_{t-1}]; \psi)}] \\
& - \mathbb{E}_{h_{t-1}, o_t} [Q(h_{t-1}, o_t, a_t; \phi)|_{h_{t-1}=\text{LSTM}(h_{t-2}, [o_{t-1}; a_{t-1}]; \psi)}] \quad (7)
\end{aligned}$$

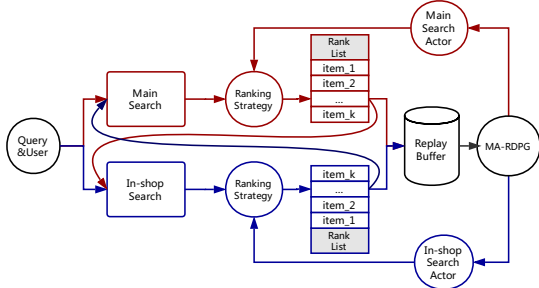
The training process is shown in Algorithm 1. We use a replay buffer [30] to store the complete trajectories to learn with minibatch update, rather than online update. At each training step, we sample an minibatch of episodes and process them in parallel to update the actor networks and the critic network respectively.

4 APPLICATION

Previous sections present a general multi-agent reinforcement learning framework that may be applicable to many joint optimization scenarios. To evaluate the proposed model, we apply it to jointly optimize the ranking strategies in two search scenarios in Taobao, which is a real-world E-commerce platform.



(a) The two search engines are optimized separately.



(b) The two systems work collaboratively in MA-RDPG.

Figure 5: Comparison of two search systems that are optimized separately or collaboratively.

Firstly, we give a brief overview of the online E-commerce platform. Then, we explain the details of how we apply our MA-RDPG to Taobao.

4.1 Search Scenarios of an E-commerce Platform

An E-commerce platform generally consists of multiple search scenarios, each of which has its own ranking strategy. In particular, we choose two important search scenarios of an E-commerce platform for this study: the main search and the in-shop search. The two search types are detailed as follows:

Main search ranks the relevant items when a user issues a query through the search box in the entrance page of the E-commerce platform. The main search returns various items from different sub-domains in the platform. The main search occupies the majority of the user traffic. In our platform, there are about 40,000 queries of main search per second. Within one day, there could be about 3.5 billion page views and 1.5 billion clicks from more than 100 million customers.

In-shop search ranks items in a certain shop when a user browses products at a shop’s page⁴. During the in-shop search, customers can search either with an input query or without any query. In one day, more than 50 million customers make shopping via in-shop search, amounting to 600 million clicks and 1.5 billion page views.

Users constantly navigate cross the two scenarios. When an user find a dress that she likes in the main search, she may go into the shop site for more similar products. When the user finds that the clothes in the shop are too limited, the user may go back to the main search for more products from other shops. Our investigation suggests that among all user shopping behavior data in Taobao, 25.46% switches from the main search to the in-shop search and 9.12% switches from the in-shop search back to the main search.

In existing models [10, 17, 24], different ranking strategies in different scenarios are independently optimized, and each strategy maximizes its own objective and ignores those of the other strategies. Figure 5(a) describes a traditional optimization method for dealing with multiple search scenarios in online platforms. The upper block in red denotes the main search engine and the lower block in blue denotes the in-shop search engine. The two search engines are optimized separately and independently.

4.2 Joint Optimization of Multi-scenario Ranking

We illustrate a solution to jointly optimizing ranking strategies in the main search and in-shop search in Figure 5(b). Instead of separately optimizing the ranking strategies in the two search scenarios, MA-RDPG employs two agents (actors) to model the two strategies collaboratively. The main search and in-shop search actors learn the weights of features in the ranking algorithms for the two scenarios respectively. The two actors collaborate in two ways: First, they have the same goal to optimize the overall performance of the system; Second, they share and broadcast messages through the communication component such that both of them have access to all historical information in different scenarios.

To be concrete, we will introduce the key concepts when MA-RDPG is applied to the scenarios.

Environment. The environment is the online E-commerce platform. Its state changes when the two agents (actors) take actions to present different ranking items. It offers rewards to the actors which also take as input the observations from the environment.

Agents. There are two agents: one is the search engine for main search and the other is that for in-shop search. At each step, one of the search engines returned a ranked list of products according to the ranking algorithm (linearly summing the features values with the feature weights). The two agents work together to maximize the overall performance, GMV, for instance.

States. As aforementioned, the states are partially observable. Agents can only receive a local observation which includes: the attributes of the customer (age, gender, purchasing power, etc.), the properties of the customer’s clicked items (price, conversion rate, sales volume, etc.), the query type and the scenario index (main or in-shop search). A 52-dimension vector is then formed to represent

⁴Some E-commerce systems such as Taobao or JingDong are the same as the real marketplaces which have many shops. Each shop sales its own products.

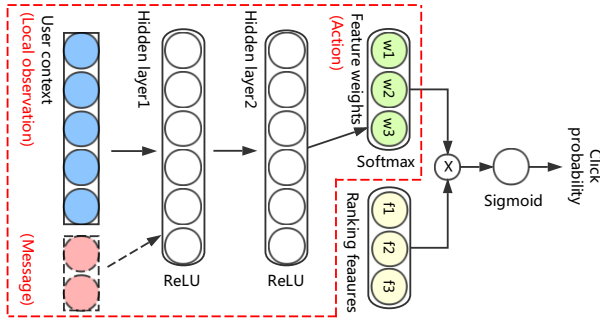


Figure 6: Actor network. The actor network in red dashed box outputs an real-valued action vector (green) for ranking given the input of local observation (blue) and message(red).

the observed information. As shown in MA-RDPG, the complete state vector are concatenation of the local observation vector and message vector which encodes historical observations and actions.

Actions. The agent needs to provide a ranking list of relevant items in response to an input query (or sometimes no query). Thus, the action of the agents is defined as the weight vector for the ranking features. To rank items, the ranking algorithm computes an inner product of the feature value vector and the weight vector. Changing an action means to change the weight vector for the ranking features. For main search, we set the actor’s action as a 7-dimension real-valued vector. For in-shop search, the action is a 3-dimension real-valued vector.

Each agent has its own policy function. The architecture of the actor network is shown in Figure 6. The actor network is a three-layer Perceptrons (MLP) with ReLU activation functions for the first two layer and softmax for the output layer. The input to the actor network is the local observation vector and the message vector. The output is the weight vector for the ranking features.

Reward. We design the rewards by considering not only purchase behaviors but also other user behaviors. In this manner, we can make full use of user feedback on the presented product list. If a purchase behavior happens, there is a positive reward that equals to the price of the bought product. If a click happens, there is a positive reward of 1. If there is no purchase nor click, a negative reward of -1 is received. If a user leaves the page without buying any product, there is a negative reward of -5 .

Table 1: Examples of Ranking Features

Scenario	Feature Name	Description
Main Search	Click Through Rate	An CTR estimation using logistic regression, considering features of users, items and their interactions
	Rating Score	Average user ratings on a certain item
	Shop Popularity	Popularity of the item shop
In-shop Search	Latest Collection	Whether an item is the latest collection or new arrivals of the shop
	Sales Volume	Sales volume of an in-shop item

5 EXPERIMENT

To evaluate the performance of our proposed MA-RDPG model, we deployed our model on Taobao to jointly optimize the main search and in-shop search.

5.1 Experiment Setting

Training Process. The flow chart of our model is shown in Figure 5(b). Our training process is based on an online learning system which consumes unbounded streams of data. Firstly, the system collects user logs in real time and provides training episodes for MA-RDPG. Secondly, the episodes are stored in a replay buffer. Thirdly, gradients are computed to update model parameters using the episodes sampled from the replay buffer. At last, a new, updated model is deployed to the online system. The process repeats. Thus, the online model is changing periodically and dynamically to capture the dynamics of user behaviors.

Parameter Setting. For each agent, the local observations is a 52-dimensional vector. The dimension of the action vector is 7 and 3 for the main and in-shop search respectively. As the communication component and critic network will take the action vectors of both actors as input, for convenience, a vector of a normalized length 10 (7+3) with zero-padding is taken as input to the LSTM and critic networks.

For the communication component, the input is a $52 + 7 + 3 = 62$ dimensional vector and the output message is a 10-dimension vector. The network structure is shown in Figure 4. In the actor network, the dimension of the input layer is $52 + 7 + 3 = 62$. The actor network was parameterized by a three-layer MLP with 32/32/7 (or 3) neurons for the first/second/third layer, respectively. The activation functions are ReLU for the first two layers and softmax for the output layer. The network structure is shown in Figure 6. The critic network has two hidden layers with 32 neurons per layer. The ReLU activation function is also used.

The reward discount factor is $\gamma = 0.9$. In our experiments, we used RMSProp for learning parameters with a learning rate of 10^{-3} and 10^{-5} for the actor and critic network respectively. We used a replay buffer size of 10^4 and the minibatch size is 100.

5.2 Baseline

The ranking algorithms in our baselines are as follows:

Empirical Weight (EW). This algorithm applies a weighted sum of the feature values with feature weights where the weights were empirically adjusted by engineering experts.

Learning to Rank (L2R). This ranking algorithm learns feature weights by a point-wise learning-to-rank network whose structure is the same as the actor network shown in Figure 6 but without message as input. The network is supervised by the user feedback of whether a click/purchase happens on an item.

The main difference among EW, L2R and MA-RDPG is the way to generate the feature weights. In MA-RDPG, feature weights are produced by the actor networks. Some typical ranking features are listed in Table 1.

On top of the algorithms, we compared MA-RDPG with three baselines that separately optimized the ranking strategies in the main search and in-shop search: 1) EW+L2R; 2) L2R+EW; 3) L2R+L2R.

Table 2: GMV gap evaluated on an online E-commerce platform. A+B means algorithm A is deployed for the main search and B for the in-shop search. The values are the relative growth ratio of GMV compared with the EW+EW setting.

day	EW + L2R			L2R + EW			L2R + L2R			MA-RDPG(ours)		
	main	in-shop	total	main	in-shop	total	main	in-shop	total	main	in-shop	total
1	0.04%	1.78%	0.58%	5.07%	-1.49%	3.04%	5.22%	0.78%	3.84%	5.37%	2.39%	4.45%
2	0.01%	1.98%	0.62%	4.96%	-0.86%	3.16%	4.82%	1.02%	3.64%	5.54%	2.53%	4.61%
3	0.08%	2.11%	0.71%	4.82%	-1.39%	2.89%	5.02%	0.89%	3.74%	5.29%	2.83%	4.53%
4	0.09%	1.89%	0.64%	5.12%	-1.07%	3.20%	5.19%	0.52%	3.74%	5.60%	2.67%	4.69%
5	-0.08%	2.24%	0.64%	4.88%	-1.15%	3.01%	4.77%	0.93%	3.58%	5.29%	2.50%	4.43%
6	0.14%	2.23%	0.79%	5.07%	-0.94%	3.21%	4.86%	0.82%	3.61%	5.59%	2.37%	4.59%
7	-0.06%	2.12%	0.62%	5.21%	-1.32%	3.19%	5.14%	1.16%	3.91%	5.30%	2.69%	4.49%
avg.	0.03%	2.05%	0.66%	5.02%	-1.17%	3.09%	5.00%	0.87%	3.72%	5.43%	2.57%	4.54%

The first algorithm indicates the one used for the main search, and the second one for the in-shop search.

5.3 Result

5.3.1 Metric. We reported the relative improvement between the compared model against the model in which EW is deployed on both scenarios (main and in-shop search), EW+EW. The metric, GMV gap, is defined as $\frac{GMV(x) - GMV(y)}{GMV(y)}$, the relative GMV growth of a model ($GMV(x)$) compared to the setting of EW+EW ($GMV(y)$). To make a fair comparison, all the algorithms run seven days in our A/B test system where 3% users are selected into the test group. The performance is measured in terms of the GMV gap in both search scenarios in these days. The performance for each single scenario is also provided as an indicator so that we can study the correlation between the two scenarios.

5.3.2 Result Analysis. The results are shown in Table 2 and we made the following observations:

First, our MA-RDPG performs much better than all the baselines which are equipped with L2R or empirical weights. In particular, MA-RDPG outperforms L2R+L2R which is a strong model currently using by Taobao, but L2R+L2R independently optimizes the ranking strategies in main search and in-shop search. It justifies that the collaboration between scenarios truly improves the overall GMV.

Second, with MA-RDPG, the GMV of in-shop search is improved significantly while the main search agent maintains comparable GMVs. The reason is that the traffic from the main search to the in-shop search is much more than that from the in-shop search to the main search (25.46% vs. 9.12%). Thus, the in-shop search agent is benefited more by receiving messages from the main search agent.

Third, the results of L2R+EW further validate our motivations that the two scenarios should cooperate with each other because improving GMV in the main search hurts that in the in-shop search.

In the lower sub-figure of Figure 7 we investigated the stability of MA-RDPG by plotting the mean performance which averages GMV gaps at the same hour within the seven days. It shows that MA-RDPG makes stable and continuous improvement.

5.3.3 Action Analysis. As aforementioned, we employed continuous actions for the agents. We thus evaluated how the actions change over time, as shown in Figure 8. Since each dimension of an

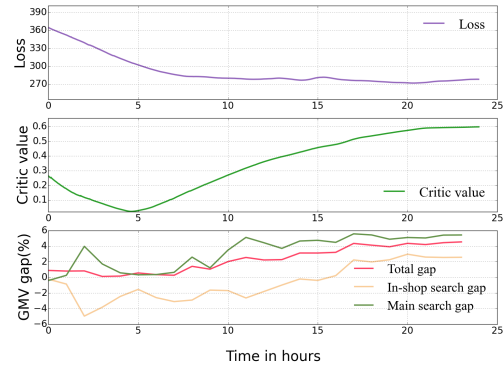


Figure 7: Upper/Middle: Learning process of the critic/actor network respectively. Lower: GMV gap against the EW+EW baseline in the online experiments.

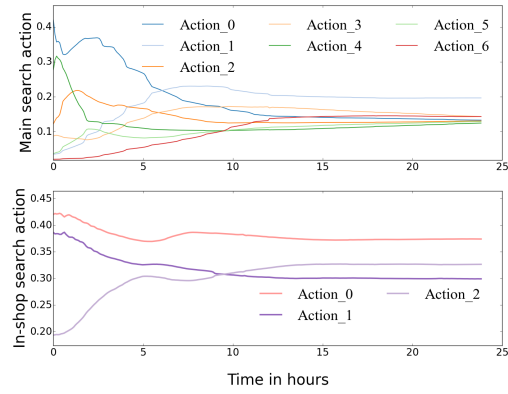


Figure 8: The change of main and in-shop search actions. Actions are averaged over the outputs of an actor network within a training batch.

action vector is real-valued, we reported the average of the action vectors in a training batch in this experiment.









dress		dress	
	Vero Moda 2017 New Dress \$123 510 Sold		Flora Dress \$35 2997 Sold
	ONLY Dress New Collection \$55 566 Sold		Jersey Knit Dress \$36 989 Sold
	Vero Moda High Waist Dress \$92 329 Sold		Retro Flouncy Dress \$36 1350 Sold
	ZARA Loose Dress \$61 322 Sold		Turtleneck Loose Dress \$52 997 Sold

Figure 9: Search result comparison. The left is by MA-RDPG and the right by L2R+L2R.

The upper sub-figure of Figure 8 depicts how the actions of the main search change over time. **Action_1** has the largest value in the main search, corresponding to the feature of Click-Through-Rate (see Table 1). This indicates that Click-Through-Rate is the most important feature, in line with the fact that CTR is known as a very important factor. **Action_6** is second largest, and it represents the weight of Shop Popularity (but not Item Popularity). However, this feature used to be a weak one in L2R, but plays a much more important role in our experiment than expected. With this feature, the main search can direct more traffic to the in-shop search by providing products from popular shops.

The change of the actions of the in-shop search is illustrated in the lower figure of Figure 8. **Action_0** is the most influential feature, which represents the weight of Sales Volume (see Table 1). More popular items seem to be bought more. Though the values of the actions varied dramatically at the early stage, they converged to stable values after about 15 hours’ training. This is accordant with the loss and critic value curves as shown in Figure 7 which shows that the critic and actor networks converged finally.

5.4 Case Study

In this subsection we further analyzed a case on how main search and in-shop search cooperate by MA-RDPG. The case illustrates how the main search helps the in-shop search, thereby targeting more future overall rewards. We simulated a scene from user log like this: a young woman with strong purchase intent clicked some items of skirt which are expensive and have low conversion rates, then she queried “dress” in the main search. The results returned by the two models are shown in Figure 9. Obviously, the results of MA-RDPG are with lower sales (small sold numbers) but with more expensive prices from more branded shops, which makes customers enter the shops with a high probability. By contrast to L2R+L2R, the main search with MA-RDPG ranks items from a global perspective in that it does not only consider its own immediate reward but also the future potential purchase during the in-shop search.

6 RELATED WORK

Ranking is a fundamental problem in many applications such as searching, recommendation, and advertising systems. A good ranking strategy can significantly improve user experience and the

performance of the applications. Learning to rank (L2R) is one typical genre of popular ranking algorithms [27, 33], and has been widely applied to E-commerce search [23], web search [1, 34], recommendation system [39, 40]. The diversity of ranking results is a common issue studied by the community, as addressed by maximal marginal relevance [7], topic representation [19], and reinforcement learning [49]. The efficiency issue is another important problem for large online platforms, which can be addressed by feature selection [14, 45, 47], cascade learning [32, 44, 46], and many other techniques.

Online large platforms generally have multiple ranking scenarios in multiple sub-domains, however, joint optimization for multi-scenario ranking is rather unexplored. We cast the problem as a fully cooperative, partially observable multi-agent reinforcement learning problem. Multi-agent reinforcement learning problem can be grouped into three categories [5]: fully cooperative, full competitive, and mixed strategies. Our task is formulated as a fully cooperative problem, which has a long history in multi-agent learning [9, 26, 36]. Due to the increased complexity of the environment, recent research efforts are paid to developing deep reinforcement learning (DRL) models. [16] combined three training schemes with DRL models to make agents collaborate with each other. Counterfactual Multi-Agent Policy Gradients [13] uses a centralised critic to estimate a global Q-function. In addition, it uses a counterfactual baseline that marginalises out the action of a single agent, to address the challenges of multi-agent credit assignment. [41] introduced a novel additive value-decomposition approach over individual agents instead of learning a shared total reward. However, agents in these three models cannot communicate with each other. Thus, the communication protocols [12] were proposed to make agents collaborate more easily. The base model of [12] is deep Q-Learning, which is not suitable for continuous action space.

Therefore, we propose our own Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) model.

7 CONCLUSION

In this paper, we present a multi-agent reinforcement learning model, MA-RDPG which employs continuous actions, deterministic policies, and recurrent message encodings. The model can optimize ranking strategies collaboratively for multi-scenario ranking problems. The model consists of a centralized critic, private actors (agents), and a communication component. Actors (agents) work collaboratively in two manners: sharing the same action-value function (the critic) that estimates the future overall rewards, and sending communication messages that encode all historical contexts. The model demonstrates advantages over baselines through online evaluation on an E-commerce platform.

The proposed model is a general framework which may be applicable to other joint ranking/optimization problems. We leave this as future work.

8 ACKNOWLEDGEMENT

This work was partly supported by the National Science Foundation of China under grant No.61272227/61332007.

REFERENCES

- [1] Leif Azzopardi and Guido Zuccon. 2016. Advances in formal models of search and search behaviour. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 1–4.
- [2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 89–96.
- [3] Christopher J Burges, Robert Ragno, and Quoc V Le. 2007. Learning to rank with nonsmooth cost functions. In *NIPS*. 193–200.
- [4] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews*, 38 (2), 2008 (2008).
- [5] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2010. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1* 310 (2010), 183–221.
- [6] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*. ACM, 129–136.
- [7] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *the 21st ACM SIGIR*. ACM, 335–336.
- [8] Jack Clark. Retrieved 28 October 2015. Google Turning Its Lucrative Web Search Over to AI Machines. *Bloomberg Business* (Retrieved 28 October 2015).
- [9] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998 (1998), 746–752.
- [10] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *ACM Conference on Recommender Systems*. 191–198.
- [11] Carl Davidson and Raymond Deneckere. 1986. Long-run competition in capacity, short-run competition in price, and the Cournot model. *The Rand Journal of Economics* (1986), 404–415.
- [12] Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. 2137–2145.
- [13] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2017. Counterfactual Multi-Agent Policy Gradients. *arXiv preprint arXiv:1705.08926* (2017).
- [14] Xiubo Geng, Tie-Yan Liu, Tao Qin, and Hang Li. 2007. Feature selection for ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 407–414.
- [15] Fredric C Gey. 1994. Inferring probability of relevance using the method of logistic regression. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 222–231.
- [16] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multiagent control using deep reinforcement learning. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2017)*.
- [17] Saurabh Gupta, Sayan Pathak, and Bivas Mitra. 2015. *Complementary Usage of Tips and Reviews for Location Recommendation in Yelp*. Springer International Publishing. 1003–1003 pages.
- [18] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. (2015).
- [19] Jiyin He, Vera Hollink, and Arjen de Vries. 2012. Combining implicit and explicit topic representations for result diversification. In *the 35th ACM SIGIR*. ACM, 851–860.
- [20] Nicolas Heess, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. 2015. Learning continuous control policies by stochastic value gradients. In *NIPS*. 2944–2952.
- [21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [22] Junling Hu, Michael P Wellman, et al. 1998. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, Vol. 98. 242–250.
- [23] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On Application of Learning to Rank for E-Commerce Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 475–484.
- [24] Krishnaram Kenthapadi, Krishnaram Kenthapadi, and Krishnaram Kenthapadi. 2017. LijAR: A System for Job Application Redistribution towards Efficient Career Marketplace. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1397–1406.
- [25] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*. 1008–1014.
- [26] Martin Lauer and Martin Riedmiller. 2000. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer.
- [27] Hang Li. 2014. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies* 7, 3 (2014), 1–121.
- [28] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 297–306.
- [29] Ping Li, Qiang Wu, and Christopher J Burges. 2008. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*. 897–904.
- [30] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [31] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, Vol. 157. 157–163.
- [32] Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. 2017. Cascade Ranking for Operational E-commerce Search. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, August 13 - 17, 2017. 1557–1565.
- [33] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
- [34] Craig Macdonald, Rodrigo LT Santos, and Iadh Ounis. 2013. The whens and hows of learning to rank for web search. *Information Retrieval* 16, 5 (2013), 584–628.
- [35] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.
- [36] Liviu Panait and Sean Luke. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems* 11, 3 (2005), 387–434.
- [37] Tao Qin, Xu-Dong Zhang, De-Sheng Wang, Tie-Yan Liu, Wei Lai, and Hang Li. 2007. Ranking with multiple hyperplanes. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 279–286.
- [38] Juan C Santamaria, Richard S Sutton, and Ashwin Ram. 1997. Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive behavior* 6, 2 (1997), 163–217.
- [39] Bichen Shi, Georgiana Ifrim, and Neil Hurley. 2016. Learning-to-rank for real-time high-precision hashtag recommendation for streaming news. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1191–1202.
- [40] Yue Shi, Martha Larson, and Alan Hanjalic. 2010. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 269–272.
- [41] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. *arXiv preprint arXiv:1706.05296* (2017).
- [42] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
- [43] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*. 1057–1063.
- [44] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–I.
- [45] Lidan Wang, Jimmy Lin, and Donald Metzler. 2010. Learning to efficiently rank. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 138–145.
- [46] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 105–114.
- [47] Lidan Wang, Donald Metzler, and Jimmy Lin. 2010. Ranking under temporal constraints. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 79–88.
- [48] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.
- [49] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2017. Adapting Markov Decision Process for Search Result Diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, August 7-11, 2017. 535–544.
- [50] Dawei Yin, Yueneng Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al. 2016. Ranking relevance in yahoo search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 323–332.