Research Article Vol. 27, No. 6 | 18 Mar 2019 | OPTICS EXPRESS 7896

Optics EXPRESS

Leveraging double-agent-based deep reinforcement learning to global optimization of elastic optical networks with enhanced survivability

XIAO LUO,[1] CHEN SHI,[2] LIQIAN WANG,[1] XUE CHEN,[1,*] YANG LI,[1] AND TAO YANG[1]

[1]*State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011, USA*
**xuechen@bupt.edu.cn*

Abstract: As the services in elastic optical networks (EONs) are bandwidth-intensive, unpredictable and dynamic, increasing network factors are emerging to affect the performance of network survivable planning and operation from network capacity to efficiency. Most of the traditional protection and restoration approaches may become before long inefficient due to the improvement of a particular network performance metric always at the expense of others. We argue that it would be more beneficial for comprehensive optimization of network performance to consider main network metrics jointly. Moreover, the highly dynamic features of EONs call for the new generation of machine learning-based solutions that are flexible and adaptable to cope with the dynamic nature of services to perform analytics. In this paper, we investigate the problem of global optimization of network performance under survivable EON environment. Specifically, a criterion, named the whole network cost-effectiveness value with survivability (WCES), is defined to measure the overall network performance by balancing the interaction among main network metrics. Then we propose a deep reinforcement learning (DRL) -based heuristic with the objective of improving overall network performance, in which two agents are utilized to provide working and protection schemes converging toward better survivable routing, modulation level and spectrum assignment (S-RMLSA) policies. Numerical results show that the proposed criterion can efficiently measure the overall network performance, and the double-agent DRL-based heuristic can greatly improve WCES while ensuring the network survivability and paying the acceptable extra consumption of request blocking probability.

1. Introduction

The rapid growth of network traffic, especially the emergence of high-rate and bandwidth-thirsty applications, are leading to huge increase in traffic volume over the Internet. In accordance with the forecasts in [1], the increase of global IP traffic will reach 3.3 ZBytes per year by 2021, which will be attained nearly threefold over the next five years. The need for scalable, cost- and bandwidth-efficient optical networks with high throughput and flexibility becomes more critical to satisfy the tremendous requirement of network capacity. Currently, thanks to the new advances in transmission technologies, such as optical orthogonal frequency division multiplexing (O-OFDM), coherent detection, and advanced digital signal processing, elastic optical networks (EONs) [2,3], also called flexi-grid networks, are the most promising solution for promoting the network throughput, which offering much finer frequency granularities, e.g., 12.5GHz or 6.25GHz. As allocating only required bandwidth for incoming traffic demands to efficient utilization of network spectrum resource, EONs

#352603 https://doi.org/10.1364/OE.27.007896
Journal © 2019 Received 22 Nov 2018; revised 24 Jan 2019; accepted 1 Feb 2019; published 4 Mar 2019

significantly facilitate the network flexibility, reconfigurability, and agility comparing to traditional wavelength division multiplexing (WDM) networks [4].

It is known that maintaining network survivability has been considered a significant attribute, since EONs carry a huge amount of information (in the order of Tb/s) and any interruption of the data flow leads to massive data loss. Therefore, adequate protection should be preplanned for EONs so that they can continue to operate under failures [5]. Meanwhile, the traditional mode of EONs planning and operation is assumed with long upgrade periods [6], which leads to the limit of network resources, including space facilities, electric power, and human resources. It is not only important but also necessary to enable the reasonable network operation schemes to handle the incremental needs and capacity of EONs until the next network upgrade. Survivable routing, modulation level and spectrum assignment (S-RMLSA) is one of fundamental problems in EONs, which aims to effectively measure the various incoming demands with the guarantee of network survivability. S-RMLSA is used to find the appropriate working and protection routes for a source-destination node pair, and allocate adaptive modulation level and suitable frequency resource to the established working and protection light-paths [7].

Previously, a large amount of literature has tried to strengthen survivability of EONs, which is measured by one or several network performance metrics, such as spectrum utilization ratio and network energy consumption (NEC) [9,10]. Even though these proposed protection schemes benefit from spectrum- or energy-efficient allocation, they may consume more other network resources, e.g., network cost and spectral efficiency, due to the absence of global optimization of main network metrics. In fact, the main network performance metrics are interrelated, such as lower modulation format level is implemented to reach farther transmission distance with lower spectral efficiency, lower NEC may lead to higher request blocking probability (RBP), and higher network survivability may cause higher spectrum resource consumption. It is difficult to evaluate the overall network performance (ONP) with any network metric separately. Thus, the global optimization of network is crucial, which involves the requirements of both network operators and service demands to provide better overall network operation than that focus on the improvement of specific or partial network capabilities [8]. Although each network criterion may be kept sub-optimal in optimization of ONP compared with the performance improvement of dedicated network criterion, the long-term network reward of EONs is much higher and it can match the network operation mechanism better.

In our previous work, we studied the global optimization of all-optical hybrid-casting in inter-datacenter EONs by defining a network performance criterion, called cost-effectiveness value (CEV), to evaluate the ONP [8]. However, only operational expenditure (OpEx) has been considered as total network cost and the effect degree of the number of served requests is treated as constant without involving the needs of network operators in CEV. To further consummate the network performance criterion and evaluate the ONP under more practical circumstance, we explore and expand the proposed CEV with enhanced network survivability for EONs. Specifically, we study the global optimization of survivable EON's performance and aim to provide 100% protection for light-paths during any single link failure. The key contributions of this paper are summarized as follows:

1) To the best of our knowledge, this is the first work on improving the overall performance of EON with shared backup path protection (SBPP). Concretely, a network performance criterion is extended based on the CEV, named the whole network cost-effectiveness value with survivability (WCES). WCES is used to measure the ONP under the scenario of single link failure by considering more practical factors from the aspect of network operators and customers. Moreover, both capital expenditure (CapEx) and OpEx are realized in network cost to make the criterion more feasible and consummate.

2) To meet the needs of both network operators and service demands as comprehensive as possible in S-RMLSA problem with OPN improvement, a novel double-agent deep reinforcement learning based survivable routing, modulation level and spectrum assignment (DA-DRL-RMLSA) approach is proposed. It leverages working agent and protection agent to find the corresponding available working and protection routing, modulation level and spectrum assignment (RMLSA) schemes, and then the reward controller is designed to find the best match working and protection RMLSA scheme combination with the maximum WCES that fed back to those two agents to update their experiences.

3) Extensive simulations are carried out to investigate the feasibility of WCES and evaluated its efficiency in ONP optimization leveraging by the proposed DA-DRL-RMLSA under dynamic network scenario. It is observed that DA-DRL-RMLSA can achieve the improvement of ONP up to 23.3% compared with benchmark algorithms.

The rest of the paper is organized as follows. Section II presents a survey on the related works and Section III describes the problem of S-RMLSA and the proposed evaluation criterion of ONP in survivable EONs. In Section IV, the framework of deep reinforcement learning is introduced. Section V presents the principle of DA-DRL-RMLSA algorithm. Simulation results are discussed in Section VI. Finally, Section VII summarizes the paper.

2. Review of related works

Since optical fibers are subject to impairments, such as being cut, providing EONs with protection is imperative. Several techniques have been considered [11] to guarantee network survivability, e.g., span restoration, preconfigured cycles (p-Cycles), 1 + 1/1:1 end-to-end path protection, and SBPP, which are mainly divided into shared protection and dedicated protection techniques.

One of major superiorities of shared protection is that network spectrum resource consumption of protection can be reduced. In [12], the service availability oriented p-Cycle protection in EONs was investigated. The authors developed a theoretical model to analyze the service availability of light-paths protected by p-Cycles, and a service availability oriented p-Cycle configuration algorithm scenario which can effectively improve service availability. Based on K shortest-path routing and first-fit spectrum assignment, a shared path-protection scheme was developed for EONs in [13] to enhance the protection efficiency. By leveraging the sharable virtualized elastic regenerator in large-scale translucent EONs, literature [14] proposed a shared-protection scheme with fallback operation when switching-over to backup route, which provided almost the same degree of survivability as dedicated protection under double-link failures. Dedicated protection techniques offer resistance against multiple link failures and allow for instantaneous recovery. In [15], the path function of the 1 + 1 protection was exchanged with the primary toggling to the backup state, while the backup becomes primary. Under the assumption that the bandwidth allocation of the protect path could be less than that of the working path, a bandwidth squeezed restoration scheme was proposed for dedicated path-protection (DPP) in EONs.

However, existing researches rarely focus on network survivability with global network performance optimization while there are several literatures about separate network performance criterion optimization in survivable EONs. The cost-efficient multilayer protection planning algorithm in IP-over-EON was proposed in [16], which can clarify the types of the spare capacity and adaptively perform spectrum sharing among them to minimize both the spectrum resources and the number of work/backup light-paths in the optical layer. In [10], an ILP model was presented for the energy-efficient survivable grooming routing and spectrum allocation (SG-RSA) problem in software-defined EONs and proposed a shared backup path grooming protection (SBPGP) algorithm to get enough protection and less resources consumption.

As the powerful interdisciplinary science combined with the mathematics, computer and biology science, machine learning (ML) recently has been successfully applied in many aspects of life, including optical communication [17]. In [18], a crosstalk estimation model based on ML was first proposed in space division multiplexing optical network, which can be used to evaluate the crosstalk during the design for the resource allocation scheme. And simulation results of the proposed scheme showed that it can improve resource utilization without increasing total connection set-up time. In particular, reinforcement learning (RL), as an important branch of ML, has made great breakthrough in optical network planning and provisioning. In the literature [19], a routing light-path establishment method based on RL for providing the service differentiation for the light-path blocking probability in all-optical WDM networks was proposed, which improves the utilization of wavelengths efficiency. To address RMLSA problem in EONs, authors in [20] demonstrate a deep reinforcement learning based self-learning RMLSA agent to cognitive and autonomous provide feasible RMLSA schemes to incoming requests.

3. Overall network performance analysis in S-RMLSA problem

In this section, we begin by providing the model of EON and the basic S-RMLSA problem considered in this paper. Then, we describe the concept of the proposed criterion, WCES, and discuss how to measure ONP by it.

3.1 Network model

The topology of EON is defined as $g(v,\varepsilon)$ where v and ε are the sets of network nodes and links, respectively. Every two adjacent nodes z and n are connected by two directed links in opposite direction, each of which corresponds to a separate fiber link, denoted by (z, n) for the one from node z to node n and (n, z) for the one from node n to node z. We assume that the frequency band of each fiber link can F be divided into adjacent frequency slots (FSs) and the bandwidth of each FS is the same. In addition, each FS is modulated with either binary phase shift keying (BPSK) or higher level modulation formats, i.e., 2^m-QAM, where the corresponding modulation level m is denoted as 1, 2, 3, and 4. By employing higher modulation level, higher spectrum efficiency can be reached while providing the same Quality of Transmission (QoT). In our model, we assume that the modulation level only affects transmission distance and the highest available modulation level is always chosen. The capacity of one FS is denoted as C_{BPSK} Gbits/s with the modulation format of BPSK. Hence, the $m \cdot C_{BPSK}$ capacity of one FS can be represented as Gbits/s in general. We assume that the architecture of EON mainly consists of bandwidth variable transponders (BV-Ts) for add-and-drop of optical signals, bandwidth variable optical cross connects (BV-OXCs) for switching/routing for pass-through optical signal, and optical amplifiers (OAs) for compensating signal attenuation.

The incoming requests are modeled as $u_i = \{s_i, des_i, C_i, b_i\}$ where i is request index, s_i and $(s_i, des_i \in v)$ des_i are the source and destination nodes, C_i denotes the capacity of request in Gbits/s, and b_i is the identifier of survivable demand, i.e., b_i equals to 1 means protection is necessary of the request, and b_i equals to 0 means the request is protection free. Further, the SBPP [21] is applied in our model, due to high efficiency of spectrum utilization. Spectrum assignment for both working path and protection path satisfies spectrum contiguity and spectrum continuity requirements.

3.2 Overall network performance evaluation with survivability

In order to estimate the ONP with survivability, the criterion of WCES is proposed by involving more detailed requirements of network operators and customers, as shown in Eq. (1),

$$WCES = \frac{WNTC \cdot WNSA}{SNC} \tag{1}$$

where WNTC is the whole network net transmission capacity, WNSA is the whole network ability for serving amount of customer demands, and SNC is survivable network cost. The exhaustive explanations of above three parts are as follows.

$$WNTC = \alpha \cdot \sum_{i \in I_p^s} \left(C_i \cdot l_i\right) + \sum_{i \in I_{np}^s} \left(C_i \cdot l_i\right) - \sum_{i \in I_{bl}} \left(C_i \cdot l_i\right) \tag{2}$$

In Eq. (2), I_p^s and I_{np}^s are the set of served requests with and without survivability, respectively. I_{bl} is the set of blocked requests. l_i is the length of the shortest light-path of the ith request. α is a factor that reflects the importance degree of network survivability, which is a constant greater than 1. WNTC is measured as the accumulated effect of each request to network capacity and reach, in which the product of network capacity and transmission reach is applied [22]. Due to both of them are main metrics measuring effectiveness of network performance that always been considered by $\sum_{i \in I_p^s} (C_i \cdot l_i)$ $\sum_{i \in I_{np}^s} (C_i \cdot l_i)$ network operators. Specially, and are used to represent the total provided network net transmission capability (NNT) by $\sum_{i \in I_{bi}} (C_i \cdot l_i)$ survivable and non-survivable requests, meanwhile, means total lost NNT. The difference of provided and lost NNT indicates the actual NNT provided by network in the time period T (that is the total duration after all possible types of requests arriving). To further adjust the impact degree of network survivability, we add the factor α onto the NNT with survivable requests, which can be adjusted depending on the requirements of network operators for network survivability.

$$WNSA = 1 + \beta \cdot \frac{\left(\left|I_s\right| - \left|I_{mean}\right|\right)}{\left|I_{mean}\right|} \tag{3}$$

In Eq. (3), I_s is the set of all served requests including both survivable and non-survivable requests. I_{mean} is the set of all served requests when the bandwidths of incoming requests are the same and equal to the average values of overall network traffic. In addition, β is the coefficient within (0, 1] @ 0.1 granularity that adjusting the impact degree of the number of served requests to ONP. WNSA is utilized to balance the impact of provided bandwidth and the number of served requests including both survivable and non-survivable, which depends on the strategies of network operators on serving amount of customer $\left(\left|I_s\right| - \left|I_{mean}\right| / \left|I_{mean}\right|\right)$ requests. is used to measure the difference rates between the number of actual served requests and its average value under ideal conditions, according to the different regulations and control strategies applied by network $\left(\left|I_s\right| \geq \left|I_{mean}\right|\right)$, operators. When it can be treated as a positive effect to network performance due to the network serves more numbers of requests under the same total network capacity, which can better match the requirement of network service providers [25]. Otherwise a negative effect to network performance is generated.

$$SNC = \cos t_{CapEx} + \cos t_{OpEx} \tag{4}$$

$$\cos t_{CapEx} = \frac{\cos t_{CapEx}^{T} \cdot \sum_{i \in I_s} t_i}{T} \tag{5}$$

$$\cos t_{OpEx} = p_u \cdot \left(\sum_{i \in I_p^s} PC_i + \sum_{i \in I_{np}^s} PC_i \right) \tag{6}$$

The SNC is generally divided into two parts, i.e., network CapEx and OpEx. $\cos t_{CapEx}^{T}$ In Eq. (4), $cost_{CapEx}$ and $cost_{OpEx}$ are the total network CapEx and OpEx during the whole network operation period, respectively. The detailed network CapEx is shown in Eq. (5). is network CapEx when the upgrade period of network is T and t_i is the duration of the ith request. We assume that CapEx is determined in network setup stage and kept constant in a network update period T. Thus, network CapEx during the whole network operation period can be converted according to the time occupying ratio. Note that there are few reports about the network components with independent function module [14], which can add new functions and expand capacity during network operation. It would be more practical if we consider variable-CapEx of network and we will address this in our future work. Equation (6) demonstrates that network OpEx is calculated by the product of the unit cost (u.c.) of energy and total NEC. Specially, p_u is the unit cost of energy, PC_i is the energy consumption of the ith request. When the ith request is survivable, the energy consumption includes both working and protection light-paths, otherwise the energy is only consumed by working light-path. The calculation of NEC is similar with our previous work [8] by obtaining traffic-independent and traffic-dependent energy consumption of main network components in optical layer. The typical values of cost and power consumption of network elements are given in Table 1. All of these values are referred from [23] and [24], in which the node degree is denoted as d and the number of add and drop ports is denoted as a.

Table 1. Typical cost and power consumption of network elements

Network element	Traffic-independent power consumption (W)	Traffic-dependent power consumption (W/Gbits)	Element cost (c.u.)
BV-OXC	150 + 85d + 50a	0	5
BV-T	120	0.18	36
OA	110	0	1

4. Deep reinforcement learning framework

In this section, the general RL is firstly presented. Then, we describe the deep Q-learning framework.

4.1 RL

RL is an important branch of ML, where an agent makes interactions with environment providing evaluative feedback to optimize its states to receive the maximal accumulative rewards. The transformation of environment in RL can usually be described as a Markov Decision Process (MDP) [26], in which state space, action set, reward function and explicit transition probability are important elements. More formally, let us assume that the interactions between the agent and environment occur at a sequence of discrete time instants t. Then, the learning problem can be formulated by defining:

a. The state space S contains environment observed by agent, where each state $s_t \in S$.

b. The action set $A(t)$ at time step t, where action $a_t \in A(t)$ can be chosen by agent.

c. The reward function $R(s_t, a_t)$ represents the feedback signal to agent after taking action a_t in state s_t.

d. The state transition probability $P_{s_t s_{t+1}}(a_t)$ represents the probability of making a transition from state s_t to s_{t+1} after implementing action a_t.

Specially, at time step t, the agent senses the state s_t and takes action a_t, then state s_{t+1} can be got depending on $P_{s_t s_t+1}(a_t)$ and agent receives a value of rewards after action finished. The aim of agent in RL is to find an optimal policy π^* by mapping states to actions to generate a sequence of rewards, i.e., $R(s_t, a_t)$, $R(s_{t+1}, a_{t+1})$, $R(s_{t+2}, a_{t+2})$, …, such that the total payoff, $R(s_t, a_t) + \gamma.R(s_{t+1}, a_{t+1}) + \gamma^2.R(s_{t+2}, a_{t+2}) + \ldots$, is as large as possible, where γ is discount factor. The cumulative discounted reward at state s_t can be expressed by the state-value function:

$$V^{\pi}(s_t) = \mathbb{E}\left[R(s_0,a_0) + \gamma \cdot R(s_1,a_1) + \gamma^2 R(s_2,a_2) + \ldots \mid \pi, s_0 = s_t, a_0 = a_t\right]; \quad (7)$$

where $\mathbb{E}$ denotes the expectation. Equation (7) can be transformed as Eq. (8),

$$V^{\pi}(s_t) = \mathbb{E}\left\{R(s_0,a_0) + \gamma \cdot \left[R(s_1,a_1) + \gamma \cdot R(s_2,a_2) + \ldots\right] \mid \pi, s_0 = s_t, a_0 = a_t\right\}; \quad (8)$$

where the expression of $R(s_1, a_1) + \gamma.R(s_2, a_2) + \ldots$ is the expected total payoff of $V^{\pi}(s_{t+1})$. Due to the Markov property, i.e., the state at the subsequent time instant is only determined by the current state, irrelevant to the former states, the value function can be rewritten as:

$$V^{\pi}(s_t) = R(s_0,a_0) + \gamma \cdot \sum_{s_0}\left[P_{s_0\pi(s_0)}(s_1) \cdot V^{\pi}(s_1)\right], s_0 = s_t; \quad (9)$$

Hence, the optimal expected total payoff $V^*(s_t)$ can be obtained by Eq. (10), which is termed as the optimal state-value function.

$$V^*(s_t) = \max_{\pi} V^{\pi}(s_t) = R(s_t,a_t) + \max_{a_t}\left\{\gamma \cdot \sum_{s_{t+1}}\left[P_{s_t s_{t+1}}(a_t) \cdot V^*(s_{t+1})\right]\right\}, \forall s_t \in \mathcal{S}; (10)$$

The optimal policy π^* follows Bellman equation can be expressed as:

$$\pi^* = \arg\max_{\pi} V^{\pi}(s_t) = \arg\max_{a_t}\left\{\gamma \cdot \sum_{s_{t+1}}\left[P_{s_t s_{t+1}}(a_t) \cdot V^*(s_{t+1})\right]\right\}, \forall s_t \in \mathcal{S}; \quad (11)$$

4.2 Deep Q-learning

When there is the large number of state-action pairs and the separate estimation of action-value function, it is impractical to do above value iteration to get the optimal action-value, i.e., $Q_t \to Q^*$, when $t \to \infty$. Thus, Q-learning, as one of most widely use strategies, is implemented to determine the optimal policy π^*. The optimal state-action function, i.e., optimal Q-function is defined as:

$$\begin{aligned} Q^*(s,a) &= \max_{\pi} \mathbb{E}\left[R(s_0,a_0) + \gamma \cdot R(s_1,a_1) + \gamma^2 R(s_2,a_2) + \ldots \mid \pi, s_0 = s, a_0 = a\right] \\ &= R(s,a) + \max_{a}\left\{\gamma \cdot \sum_{s'}\left[P_{ss'}(a) \cdot V^*(s')\right]\right\} \quad (12) \\ &= R(s,a) + \gamma \cdot \max_{a'} Q(s',a'), \forall s, s' \in \mathcal{S}; \end{aligned}$$

Here, the objective is changed to find the proper Q-function instead of finding the optimal policy. Usually, Q-function is obtained in a recursive manner, so that the Q-function is updated as:

$$Q_{t+1}(s,a) = Q_t(s,a) + \eta \cdot \left\{R(s,a) + \mathbb{E}\left[\max_{a} Q_t(s',a') - Q_t(s,a)\right]\right\}; \quad (13)$$

where η is the learning rate. By leveraging neural network with weight θ_j, i.e., a kind of nonlinear function approximator, the basic Q-table in common RL is substituted with Q-network in deep Q-learning. In addition, the experience replay and iterative update are two main techniques to address the instabilities in Q-network [27]. The deep Q-network can be trained by updating θ_j at every iteration j to reduce the error between input and output in Q-function, where the optimal target values are $R(s,a) + \gamma \cdot \max_{a'} Q(s',a')$ replaced by

$R(s,a)+\gamma\cdot\max_{a'}Q(s',a',\theta_j^-)$ Weight is obtained from previous iteration by experience replay operation. The loss function $Loss(\theta_j)$ that changes at each iteration j can be expressed as,

$$Loss(\theta_j)=\mathbb{E}\left\{\left[R(s,a)+\gamma\cdot\max_{a'}Q(s',a',\theta_j^-)-Q(s,a,\theta_j)\right]^2\right\}; \tag{14}$$

By applying the DRL framework into S-RMLSA, we implement the global optimization of ONP with enhanced survivability. The detailed heuristic is described in the following section.

5. Double-agent based S-RMLSA algorithm

In this section, we propose the DA-DRL-RMLSA to reach the higher ONP by maximizing the value of WCES, in which the working and the backup RMLSA agents are achieved.

5.1 Basic elements in DA-DRL-RMLSA

1) Network state: we define that the current network state is jointly determined by the states of all FS's in the network. In particular, the network state $s(t)$ at time slot t is defined as,

$$s(t)=\left\{\left[\mathcal{B}_{d,e}^t\right]\right\}, d=1,2,...,|\mathcal{E}|; e=1,2,...,\mathcal{F}; \tag{15}$$

where $\mathcal{B}_{d,e}^t$ denotes the state of the eth FS in link d at time slot t, e.g., $\mathcal{B}_{7,2}^t$ = 1 means the second FS of the seventh fiber link is available to be allocated at time slot t, $\mathcal{B}_{5,31}^t$ and = 0 means the thirty-first FS of the fifth fiber link is occupied at time slot t. The number of possible network states is very large that bringing about high dimensionality of data. Fortunately, deep Q-network is efficient enough to learn directly from high-dimensional inputs, thus it is proper to be used in S-RMLSA formulation.

2) Action: In S-RMLSA, the working and protection agents have to decide the routing path, select modulation level, and allocate continuous FS resources in working and protection dimension for each incoming request, respectively. The current sets of working action (WA) and protection action (PA) are defined in Eqs. (16) and (17), respectively.

$$WA(t)=\left\{\left[a_{W,k_1}(t)\right]\right\}, k_1=1,2,...,K_1; \tag{16}$$

$$PA(t)=\left\{\left[a_{P,k_2}(t)\right]\right\}, k_2=1,2,...,K_2; \tag{17}$$

In Eqs. (16) and (17), $WA(t)$ and $PA(t)$ denote the corresponding candidate working and protection action sets. $a_{W,k_1}(t)$ and $a_{P,k_2}(t)$ represent the k_1th and k_2th working $a_{W,k_1}(t)=\{l_{P,k_1},m_{W,k_1},[f_{st,k_1},f_{en,k_1}]\}$, action and l_{Pk_1} protection action, respectively. For each working action, we l_{Pk_1} fully define it as in which m_{W,k_1}, f_{st,k_1} is the f_{em,k_1} routing light-path selected from the set of $a_{P,k_1}(t)=\{l_{P,k_1},m_{P,k_2},[f_{st,k_2},f_{en,k_2}]\}$, candidate working routing light-paths, is the l_{P,k_2} corresponding modulation level based on and request capacity. and are the indexes of start and end occupied FS for working routing light-path. m_{P,k_2} Similarly, each protection action is also defined l_{P,k_2} as is the routing f_{st,k_2} f_{en,k_2} light-path selected from the set of candidate protection routing light-paths, which the calculation of candidate protection routing light-paths is based on the corresponding working routing paths to avoid link sharing. is the corresponding modulation level based on and request capacity. and are the indexes of start and end occupied FS for protection routing light-path. K_1 and K_2 are the total number of candidate working and protection routing light-paths, respectively. We also define an action set $A(t) = \{WA(t), PA(t)\}$ to store the whole working and protection actions at time slot t.

3) Reward function: Due to the objective of S-RMLSA is to optimize ONP, the reward function is defined to maximize the value of WCES that defined in Eq. (1). Thus, reward function can be denoted as,

$$r(t) = \frac{WNTC \cdot WNSA}{SNC} \tag{18}$$

The feature of immediate reward $r(t)$ that measures the accumulative effect of previous and current actions to network performance, providing an efficient way to improve ONP. We have stored the information of all served and blocked incoming requests before current time slot, so that they can be used to calculate the immediate reward.

5.2 Training phase

Generally, the working agent selects and executes the working RMLSA action at current network state, and then the network state is updated. Next the protection agent selects and executes the protection RMLSA action based on the updated network state, and the immediate reward can be got from network environment. Then the reward controller sends the future reward for current time instant, and the Q-network is trained with the obtained action-value pairs. The full algorithm for training deep Q-network is presented in Algorithm 1.

Algorithm 1 Double-agent based deep reinforcement learning algorithm for S-RMLSA

```
Initialize Q-network with random weight $\theta$;
Initialize target Q-network with weight $\theta^{-} = \theta$;
Initialize $\gamma$ and $\varepsilon$;
for episode = 1 to N do
    Initialize the beginning state $s$;
    for t = 1 to $T$ do
    Select a random probability $P$;
    if $P \geq \varepsilon$ then
    A(t) = arg max_A Q($s_t$, $A$, $\theta$);
    Otherwise randomly select a working and protection actions to form A(t);
    Execute A(t) in EONs;
    Protection agent observe the reward r(t) and next state $s_{t+1}$;
    Store $E_t^W, E_t^P$ experience into replay memory;
    Get mini-batch of sample experiences from replay memory
    Perform gradient descent step with respect to $\theta$;
    Update $\theta^{-}$ according to Eq. (14);
    end for
end for
```

Inside the deep Q-network of S-RMLSA problem, to learn the weight θ such that the outputs $Q(s, a, \theta)$ best approximate to $Q^*(s, a)$, the training data is designed as input data set x

= {[s(t), A(t)]|t>0} and corresponding target outputs y = { Q*[s(t), A(t)]|t>0}where A(t) is the set of WA(t) and PA(t). The input data can be obtained from the experience replay memory, where both agents record the observed interaction experience $E_t^W = \left\{s(t), a_{W,k_1}(t), s(t+t_{tr})\right\}$ $E_t^P = \left\{s(t+t_{tr}), a_{P,k_2}(t+t_{tr}), s(t+1)\right\}$ together, i.e., and into $M = \left\{\left(E_1^W + E_1^P\right), \left(E_2^W, E_2^P\right), \ldots, \left(E_t^W + E_t^P\right)\right\}$ experience replay memory at each time step t, where t_{tr} is the $r\left[s(t), A(t)\right] + \gamma \cdot \max_{A(t)'} Q\left[s(t)', A(t)', \theta^-\right]$ intermediate time translating from working RMLSA to protection RMLSA. We use the approximate value to replace Q*[s(t), A(t)]. In particular, Q[s(t)', A(t)', θ^-] is the output of the target Q-network with parameter θ^-, which has the same architecture of action-value $\left(E_t^W + E_t^P\right)$ Q-network. The training data of target Q-network consists of the corresponding s(t + 1) from interaction experience, i.e., After collecting training data, two agents learn parameter θ by training Q-network to minimize the loss between actual Q-value and target Q-value, which the loss function is mentioned in Eq. (14). The whole process of double-agent training at each time step is shown in Fig. 1. We apply the ε-greedy policy [27] in reward controller to balance the exploration and exploitation of target Q-network training, in which the largest Q-values are selected with a probability of (1-ε), and the random Q-values are selected with a probability of ε.

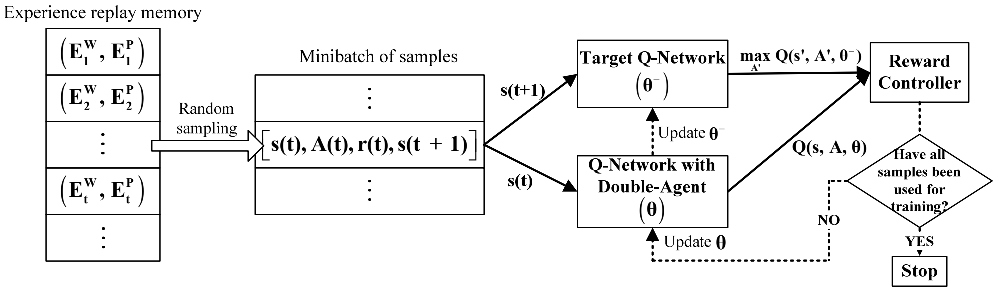

Fig. 1. Process of agent training at each time step t.

The exact structure of Q-network is demonstrated schematically in Fig. 2. Note that the small matrices and vectors in Fig. 2 are only used for the indication of functionality, and the actual dimensions of them are set accordingly in implementation. The first layer is used to incorporate input network state and the information of request by convolution and then applies a rectifier nonlinearity activation function to the results of convolution operation, which can be treated as the preprocessing of input data. The second layer convolves the output of the first layer with 16 filters of 4 × 4 with stride 2, where stride is the number of points skipped across the layer. A rectifier nonlinearity activation function is also applied to the results of convolution operation. Concretely, 16 filters are used to convolute the input data (that is the output of the first layer) with the fixed size, 4 × 4, to carry out feature extraction. Then 16 characteristics are obtained, which are treated as the output of the second layer. The third layer convolves the output of the second layer and the updated network state (that is the network state after applying working RMLSA scheme) with 32 filters of 2 × 2 with stride 1, and followed the same rectifier nonlinearity activation function. Specifically, the FS occupation matrix convolves with the corresponding 32 characteristics extracted from the input of the third layer to get the characteristics containing current network state. The fourth layer is fully connected of 128 units, followed by rectifier nonlinearity as well. The final output layer, i.e., the fifth layer, linearly full connects with the fourth layer, which outputs the corresponding estimated Q(s, A, θ)for each action set A.

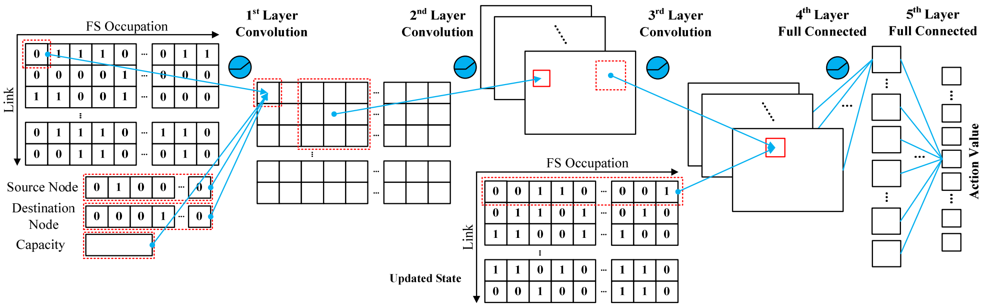

Fig. 2. The structure of Q-network.

6. Performance evaluation

In this section, numerical results for S-RMLSA problem with global optimization are presented. We investigate the algorithm performance of proposed DA-DRL-RMLSA under training phase and dynamic network scenario. Moreover, the impact of adjustable parameters in WCES is analyzed.

6.1 Simulation results and discussions in training phase

In the simulations, we consider 14-node NSFNET network topology [8] and all the light-paths (both working and protection) are set up all-optically without O/E/O conversions. The EON supports four modulation formats, i.e., BPSK, QPSK, 8-QAM, and 16-QAM, and their spectrum efficiencies and maximum transmission reaches are the same as those in [28]. All training requests and the light-path candidates for each s-d pair are generated in advance based on the given network topology. The proportion of protection and protection-free request is set as 1:1. Furthermore, we use TensorFlow 0.12.1 with Python 2.7 on Ubuntu 12.04 LTS in our simulations to implement deep reinforcement learning. Table 2 summarizes the main simulation parameter settings. The energy cost is assumed as 7.22 Cents per KiloWatthour (kWh) for industrial customers in the United States in 2017 [29]. The normalized cost value is $1.44*10^{-3}$ c.u./kWh.

Figure 3(a) and 3(b) demonstrate algorithm performance under different values of its main parameters, i.e., learning rate and mini-batch size. The dotted lines with the same colors as the corresponding learning rates are the mean values of WCES during the whole training episode in Fig. 3(a), and it shows that the values of average WCES increase with learning rate increasing, until the value of learning rate is 10^{-4} (here, the size of mini-batch is set as 8). Then the values of average WCES decrease when the learning rate further increases. This is mainly because that the impact of experience to train agents decreases with learning rate decrease, and the impact of experience is increasing to prevent agents to explore better solutions with its sustained growth. Moreover, the size of mini-batch affects the performance of algorithm as well (here, the value of learning rate is 10^{-4}), shown in Fig. 3(b). The changing trend of WCES under different mini-batch sizes is similar to that under different learning rates. This is because that the fewer samples can be used to train agents when mini-batch size decreases, and the larger number of samples lead to longer time of convergence when mini-batch size increases. Thus, the fittest values of learning rate and mini-batch size in S-RMLSA problem can be determined as 10^{-4} and 8 for the rest of simulation.

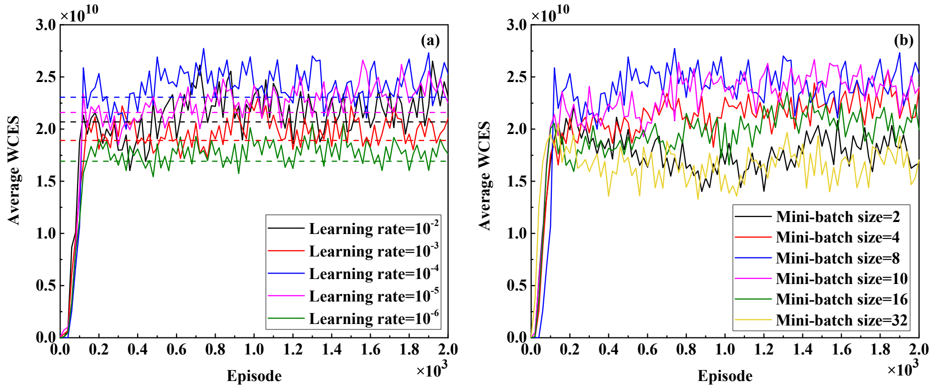

Fig. 3. Convergence performance with different (a) learning rates and (b) the sizes of mini-batch.

Table 2. Main simulation parameter settings

Bandwidth of a FS	12.5GHz
C_{BPSK}	12.5Gbtis/s
Number of training requests	20,000
Range of request capacity	50-200Gbits/s @1 Gbits/s granularity
Number of lightpath candidates for each s-d pair	4
Experience replay memory size	1000
Update frequency	4
Discount factor	0.99
Initial exploration	1
Final exploration	0.1
Target network update rate	0.001

6.2 Simulation results and discussions in dynamic network scenario

In this section, we investigate survivable EON provisioning with dynamic traffic, i.e., the requests are time variant, which can arrive and leave on the fly during network operation. To emulate practical network provisioning, requests arrive according to a Poisson process with the average request arrival rate of λ, and the lifetime of each request follows negative exponential distribution with mean value of μ. Hence, the traffic load can be measured by λ/μ in Erlangs. The parameters of proposed DA-DRL-RMLSA are set the same as training phase. The simulation network topology is also 14-node NSFNET and the number of FS's in each fiber link is assumed as 358. In addition, if we cannot provide a feasible S-RMLSA scheme (that contains both working RMLSA and protection RMLSA solutions) for a request, it is treated as blocked.

For performance comparison, we select two spectrum-efficient benchmark algorithms with P-cycle protection. The maximum-independent-set based failure-independent path protection P-cycle design (MIS-FIPP) [30] and genetic P-cycle combination protection strategy (GPCPS) [31] are implemented in our simulation environment by C + + programming directly. All simulations of these two benchmark algorithms are completed on a computer with 4.00 GB RAM, i5-4590 CPU and 3.30GHz Inter Core.

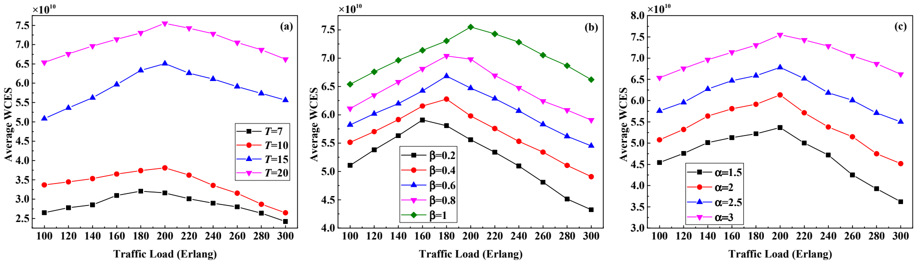

Fig. 4. Average values of WCES under different values of (a) T, (b) β and (c) α.

We evaluate the impact of parameters in WCES to ONP. The values of average WCES under different network upgrade periods T, impact degree coefficient β and factor α are shown in Figs. 4(a)–4(c), respectively. The coefficient β is the impact degree of the number of served requests to ONP, which is defined in Eq. (3) of Section 3.2, and the factor α is the importance degree of network survivability, which is defined in Eq. (2) of Section 3.2. We can observe that the average values of WCES increase with the value of T increasing from Fig. 4(a). This is because longer network upgrade period leads to lower CapEx per time unit, and further decreases the corrected network CapEx in the whole network operation time period. It is interesting to note that the change rate of average WCES is increase with enlarging the value of T and the existence of peak value. This is mainly because that the larger T leads to the faster change of the whole network CapEx. There will be more possibility to find the best match among network metrics when the value of T is larger under the same traffic load. In addition, the whole network operation time period becomes larger with the traffic load increase, the optimization degree of network cost gets lower, which further results in the value of average WCES decreases after reaching the peak value. In Fig. 4(b), we can observe that the average WCES increases with coefficient β increasing, which reflects the larger number of requests leading to the better ONP. Moreover, the value of peak also gets larger due to the number of served requests is an important factor to network performance for network operators. Figure 4(c) depicts the values of average WCES under different α. We observe that ONP becomes better when increasing the value of α. This is due to the fact that the impact degree of survivable requests increasing under the same traffic load. Moreover, the disparity among values of average WCES under different α becomes larger with the traffic load increasing. This is mainly because the number of successful served survivable requests gets larger with increasing of traffic load, which further expands the disparity of different WCES values.

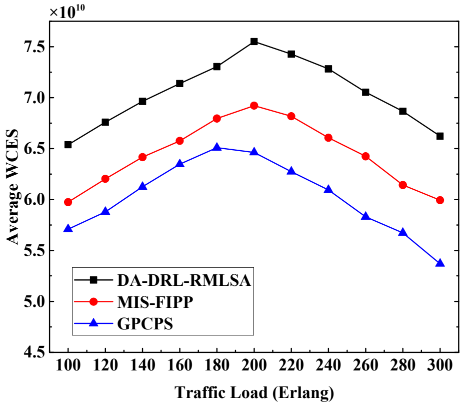

Fig. 5. Average values of WCES under different traffic loads.

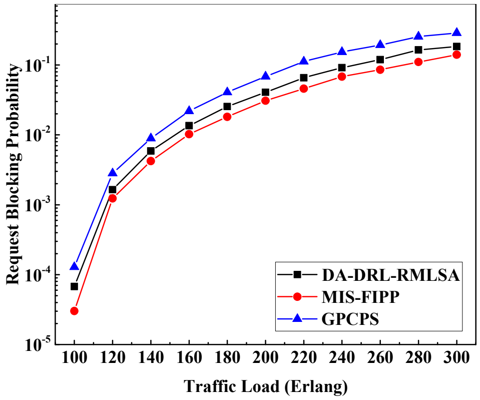

Fig. 6. Request blocking probabilities under different traffic loads.

The ONP is also evaluated by the values of average WCES provided by DA-DRL-RMLSA and benchmark algorithms, shown in Fig. 5. It is demonstrated that the proposed DA-DRL-RMLSA can provide the best ONP, i.e., the largest average WCES values, comparing with benchmark algorithms. The maximum improvement of WCES is about 23.3% under different traffic loads, compared to the results provided by GPCPS. This is because that DA-DRL-RMLSA considers the impact of future possible S-RMLSA schemes to the current S-RMLSA scheme in both working agent and protection agent to achieve a better global optimization. We also observe that the value of WCES of DA-DRL-RMLSA, MIS-FIPP and GPCPS reaches peak when traffic load is around 200, 200 and 180 Erlang, respectively. This result demonstrates the efficiency of the proposed WCES. The peak values of these three algorithms indicate that ONP reaches optimal. Under these circumstances, all of main network performance metrics are the most balanced and in relatively sub-optimal, instead of some of metrics are optimal while others are in low performances. Moreover, the peak WCES value of DA-DRL-RMLSA is larger than benchmark algorithms, which means it can efficiently coordinate WNTC, WNSA and SNC to make the network serving more heavy traffic as well as better ONP.

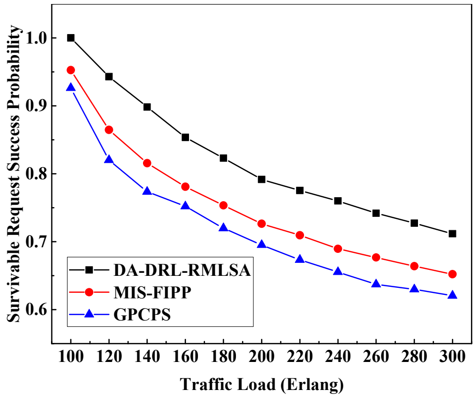

Fig. 7. Survivable request success probabilities under different traffic loads.

To evaluate the effect of optimizing ONP to main network metrics in dynamic scenarios, we simulate the RBP and survivable request success probability (SRSP) under different traffic loads in Figs. 6 and 7. It is obvious that the RBPs of MIS-FIPP keep the minimum and the RBPs of DA-DRL-RMLSA stay sub-minimum with traffic load increasing, shown in Fig. 6. The disparity of RBPs between MIS-FIPP and DA-DRL-RMLSA is about 22.7%-30.1%. This is mainly because the protection RMLSA solutions with P-cycle of MIS-FIPP efficiently increase spectrum utilization ratio and further improve service ability of network to serve more requests, while DA-DRL-RMLSA pays more attention on balancing WNTC, WNSA, and SNC leading the sub-optimal results in RBPs. Figure 7 depicts that DA-DRL-RMLSA can provide the best SRSP under different traffic loads, which reflects the optimization process of it can solve more feasible protection RMLSA schemes to increase the request survivability. The improvement of SRSP of DA-DRL-RMLSA is about 13.5%-16.4% compared to benchmark algorithms.

7. Conclusions

In this paper, the global optimization of ONP with enhanced network survivability is investigated. WNTC, WNSA, and SNC are treated as three main factors affecting ONP. A new criterion of survivable cost-effectiveness is defined by capturing above factors, which could be an effective metric to measure ONP for EONs. Moreover, a DRL-based S-RMLSA heuristic with two agents is proposed that gives a new research direction for designing disaster-resilient and improving ONP-efficient of EONs. Working agent and protection agent of DA-DRL-RMLSA explore feasible working RMLSA and protection RMLSA schemes respectively, which are combined by reward controller with the objective of maximizing WCES. Numerical results show the availability and efficiency of WCES. And they also demonstrated that the proposed DA-DRL-RMLSA can greatly optimize ONP while ensuring survivability of EONs against single-link failure. The maximum improvement of ONP is 23.3% comparing to benchmark algorithms with acceptable RBP increase.

Funding

National Natural Science Foundation of China (NSFC) (No. 61871415).

References

1. H. Takeshita, T. Oguma, S. Fujisawa, Y. Suzuki, B. Yatabe, A. Tajima, H. Hasegawa, and K. Sato, “Enhancing protection performance by effectively using spectral slots of impairment-aware elastic optical networks [invited],” J. Opt. Commun. Netw. **10**(1), A91–A101 (2018).
2. M. Jinno, “Elastic Optical Networking: Roles and Benefits in Beyond 100-Gb/s Era,” J. Lightwave Technol. **35**(5), 1116–1124 (2017).
3. M. Y. Namaad, A. G. Rahbar, and B. Alizadeh, “Adaptive modulation and flexible resource allocation in space-division- multiplexed elastic optical networks,” J. Opt. Commun. Netw. **10**(3), 240–251 (2018).
4. B. C. Chatterjee, N. Sarma, and E. Oki, “Routing and Spectrum Allocation in Elastic Optical Networks: A Tutorial,” IEEE Commun. Surv. Tut. **17**(3), 1776–1800 (2015).
5. X. Chen, M. Tornatore, S. Zhu, F. Ji, W. Zhou, C. Chen, D. Hu, L. Jiang, and Z. Zhu, “Flexible Availability-Aware Differentiated Protection in Software-Defined Elastic Optical Networks,” J. Lightwave Technol. **33**(18), 3872–3882 (2015).
6. P. Papanikolaou, K. Christodoulopoulos, and E. Varvarigos, “Optimization techniques for incremental planning of multilayer elastic optical networks,” J. Opt. Commun. Netw. **10**(3), 183–194 (2018).
7. X. Li, S. Huang, S. Yin, B. Guo, Y. Zhao, J. Zhang, M. Zhang, and W. Gu, “Shared end-to-content backup path protection in k-node (edge) content connected elastic optical datacenter networks,” Opt. Express **24**(9), 9446–9464 (2016).
8. X. Luo, C. Shi, X. Chen, L. Wang, and T. Yang, “Global Optimization of All-Optical Hybrid-Casting in Inter-Datacenter Elastic Optical Networks,” IEEE Access **6**, 36530–36543 (2018).
9. F. Ji, X. Chen, W. Lu, J. J. P. C. Rodrigues, and Z. Zhu, “Dynamic p-cycle configuration in spectrum-sliced elastic optical networks,” *in Proceedings of IEEE Global Communications Conference* (*CLOBECOM 2013*), 2170–2175.
10. J. Wu, Z. Ning, and L. Guo, “Energy-Efficient Survivable Grooming in Software-Defined Elastic Optical Networks,” IEEE Access **5**, 6454–6463 (2017).
11. G. Shen, H. Guo, and S. K. Bose, “Survivable elastic optical networks: survey and perspective (invited),” Photonic Netw. Commun. **31**(1), 71–87 (2016).
12. X. Chen, F. Ji, and Z. Zhu, “Service availability oriented p-cycle protection design in elastic optical networks,” J. Opt. Commun. Netw. **6**(10), 901–910 (2014).
13. X. Shao, Y. Yeo, Z. Xu, X. Cheng, and L. Zhou, “Shared-path protection in OFDM-based optical networks with elastic bandwidth allocation,” *in Proceedings of Optical Fiber Communication Conference and Exposition* (*OFC 2012*), paper 1–3.

14. M. Jinno, T. Takagi, and Y. Uemura, "Enhanced survivability of translucent elastic optical network employing shared protection with fallback," *in Proceedings of Optical Fiber Communication Conference and Exposition* (*OFC 2017*), paper Th3K.1.
15. S. Ba, B. C. Chatterjee, and E. Oki, "Defragmentation Scheme Based on Exchanging Primary and Backup Paths in 1+1 Path Protected Elastic Optical Networks," IEEE ACM T. Network. **25**(3), 1717–1731 (2017).
16. W. Lu, X. Yin, X. Cheng, and Z. Zhu, "On cost-efficient integrated multilayer protection planning in IP-over-EONs," J. Lightwave Technol. **36**(10), 2037–2048 (2018).
17. D. Wang, M. Zhang, J. Li, Z. Li, J. Li, C. Song, and X. Chen, "Intelligent constellation diagram analyzer using convolutional neural network-based deep learning," Opt. Express **25**(15), 17150–17166 (2017).
18. Q. Yao, H. Yang, R. Zhu, A. Yu, W. Bai, Y. Tan, J. Zhang, and H. Xiao, "Core, Mode, and Spectrum Assignment Based on Machine Learning in Space Division Multiplexing Elastic Optical Networks," IEEE Access **6**, 15898–15907 (2018).
19. I. Koyanagi, T. Tachibana, and K. Sugimoto, "A reinforcement learning-based lightpath establishment for service differentiation in all-optical WDM networks," *in Proceedings of IEEE Global Communications Conference* (*CLOBECOM 2009*), paper 1–6.
20. X. Chen, J. Guo, Z. Zhu, R. Proietti, A. Castro, and S. J. B. Yoo, "Deep-RMSA: A Deep-Reinforcement-Learning Routing, Modulation and Spectrum Assignment Agent for Elastic Optical Networks," *in Proceedings of Optical Fiber Communication Conference and Exposition* (*OFC 2018*), paper W4F.2.
21. A. Cai, J. Guo, R. Lin, G. Shen, and M. Zukerman, "Multicast Routing and Distance-Adaptive Spectrum Allocation in Elastic Optical Networks With Shared Protection," J. Lightwave Technol. **34**(17), 4076–4088 (2016).
22. M. Eiselt, B. T. Teipen, K. Grobe, A. Autenrieth, and J. P. Elbers, "Programmable modulation for high-capacity networks," *in Proceedings of European Conference and Exhibition on Optical Communication* (*ECOC 2011*), paper 1–3.
23. R. Huelsermann, M. Gunkel, C. Meusburger, and D. A. Schupke, "Cost modeling and evaluation of capital expenditures in optical multilayer networks," J. Opt. Netw. **7**(9), 814–833 (2008).
24. A. Fallahpour, H. Beyranvand, and J. A. Salehi, "Energy-Efficient Manycast Routing and Spectrum Assignment in Elastic Optical Networks for Cloud Computing Environment," J. Lightwave Technol. **33**(19), 4008–4018 (2015).
25. I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an Efficient C-RAN Optical Fronthaul for the Future Networks: A Tutorial on Technologies, Requirements, Challenges, and Solutions," IEEE Commun. Surv. Tut. **20**(1), 708–769 (2018).
26. Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover Control in Wireless Systems via Asynchronous Multi-User Deep Reinforcement Learning," IEEE Internet Things **5**(6), 4296–4307.
27. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," Nature **518**(7540), 529–533 (2015).
28. M. Moharrami, A. Fallahpour, H. Beyranvand, and J. A. Salehi, "Resource Allocation and Multicast Routing on Elastic Optical Networks," IEEE Trans. Commun. **65**(5), 2101–2113 (2017).
29. U.S. Energy Information Administration (EIA), https://www.eia.gov/.
30. X. Chen, S. Zhu, L. Jiang, and Z. Zhu, "On Spectrum Efficient Failure-Independent Path Protection *p*-Cycle Design in Elastic Optical Networks," J. Lightwave Technol. **33**(17), 3719–3729 (2015).
31. X. Guo, J. Huang, H. Liu, and Y. Chen, "Efficient P-cycle combination protection strategy based on improved genetic algorithm in elastic optical networks," IET Optoelectron. **12**(2), 73–79 (2018).