Interpreting Multi-Objective Reinforcement Learning for Routing and Wavelength Assignment in Optical Networks

SAM NALLAPERUMA[1,*], ZELIN GAN[1], JOSH NEVIN[2], MYKYTA SHEVCHENKO[3], AND SEB J. SAVORY[1]

[1] Fibre Optic Communication Systems Laboratory (FOCSLab), Electrical Engineering Division, Department of Engineering, University of Cambridge, 9 JJ Thomson Avenue, Cambridge CB3 0FA, U.K.
[2] CloudNC, 1 Norton Folgate, Spitalfields, London E1 6DB, U.K.
[3] Department of Electronic & Electrical Engineering, University College London (UCL), Torrington Place, London WC1E 7JE, U.K.
[*] Corresponding author: snn26@cam.ac.uk

Compiled July 17, 2023

Performance optimization literature in optical networks predominantly consists of single objective optimization studies while often in practice multiple performance goals are to be met. This study addresses this issue with a generalized reinforcement learning (RL) model for parameter optimization in optical networks in the presence of multiple performance goals. Using this generic model, two multi-objective variants of a classical optimization problem in optical network operation, routing and wavelength assignment (RWA) are derived and solved to near optimality. The allocated route and wavelength for each demand are optimized with respect to the number of accepted services, the number of transmitters and network availability. The resultant approximated Pareto front provides a set of solutions from which network operators can make decisions based on their preferences for particular objectives. These results contribute to the understanding into the relationships between different network parameters and performance metrics which would be beneficial in future network design and growth. Moreover, benchmarking results against the state-of-the-art RWA heuristics suggest the applicability of RL in dynamic settings under changing traffic and generalizability for unseen traffic.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

It is vital to understand the effects from network parameters in order to improve the network performance given the flexibility and costs associated with setting these parameters. Several studies have been conducted using exact [1–6] and stochastic [7–13] methods considering different network parameters and performance metrics. Most of these studies have considered a single objective or performance goal, for example, maximizing throughput, minimizing latency, minimizing cost or maximizing resilience, etc. However, optimizing a single objective may negatively impact other metrics that are important in real world applications [14]. While it is of research interest to understand these performance goals individually, the simultaneous optimization of multiple objectives is critical for the design and operation of real-world optical networks.

As case studies, we consider two multi-objective variants of the routing and wavelength assignment (RWA), a well known optimization problem in optical network literature. RWA is proven to be a $\mathcal{NP}-$hard problem [15], meaning no exact approach exists that guarantees an optimal solution in polynomial time. In the literature, several problem formulations have been proposed to solve RWA in optical networks [16, 17].

Recently, in machine learning and operational research fields, the potential of reinforcement learning (RL) in multi-objective optimization has been investigated. In the context of RL, multi-objective optimization can be realized as multi-policy optimization, where the preference of multiple objectives are not known in advance. Multi-policy methods have multiple policies representing different preference functions and has shown potential in various applications in recent years [18–20].

This study presents a novel approach for solving the multi-objective RWA problem in optical networks under dynamic traffic conditions. To the best of our knowledge, this is the first study to tackle this challenging problem, which involves optimizing network throughput, cost, and resilience simultaneously. Traditional meta-heuristic [21–23] and exact [24] multi-objective optimization approaches are not practical in this setting due to their high running times, which make them impractical for online servicing of new traffic requests. In contrast, our approach utilizes RL to rapidly provide optimal RWA solutions, making it

a practical and efficient solution for this problem. To this end, this study offers a significant contribution to the field of optical network design and operation.

The results showcase the optimized solutions considering different preference functions with certain preference weights for each objective. These can be useful in decision making processes for network design and operation which consider multiple often conflicting goals. As shown in benchmarking results, RL approach was able to learn offline and run online swiftly in the same order of runtime and solution quality as industrially deployed RWA heuristics such as k-shortest path first fit (kSP-FF). Moreover, as shown in the results, RL agent trained on uniform traffic was able to be generalized for unseen population based traffic (Eq. 12). These results suggest the potential of RL to be applied under dynamic settings with changing traffic.

The organization of the remainder of this paper is as follows. Section 2 describes the optical network physical layer model and the dynamic RWA problem underpinning the proposed framework. In Section 3, we outline the proposed multi-objective RL framework, followed by Section 4 describing the considered simulation set up. Section 5 presents simulation results for the bi-objective case, which is followed by Section 6 in which we generalize the results for multi-objective optimization and constrained optimization. In Section 7, we provide interpretations of the learned RL policies. Section 8 extends the simulations for non-uniform traffic. Finally, in Section 9 we present our concluding remarks.

2. PRELIMINARIES

A. Physical Layer of Optical Networks

We make the simplifying assumption of transmission at the Shannon rate by assigning the point-to-point capacity between a given source and a destination node per light-path as the theoretical upper-bound taken at the optimum launch power [11]. Thus, the capacity per light-path is defined as follows

$$\mathsf{C}_{p_j} = 2R_S \log_2\left(1 + \mathrm{SNR}_{p_j}\right) \quad \textbf{(1)}$$

where R_S is the symbol rate, and SNR_{p_j} stands for the signal-to-noise ratio (SNR) at the end of the p_j-light-path [25]. System parameters are chosen as a carrier wavelength of 1550 nm, a symbol rate (R_S) of 100 GBd, a channel spacing of 100 GHz, the number of channels as 100, a span length of 100 km[1], the EDFA noise figure as 4.5 dB and the attenuation coefficient (α_{dB}) as 0.2 dB/km.

B. Dynamic Routing and Wavelength Assignment Problem

This study considers the dynamic RWA problem in which lightpath requests arrive and expire over time. Dynamic RWA is not commonly formulated as an integer linear programming problem like the static RWA. Instead, dynamic RWA is solved using online algorithms [13] and heuristics [26], which adapt to the changing network conditions. A general outline of the dynamic RWA problem using arrival and expiration events can be described as follows:

Objective:
Maximize network throughput over time.

[1]the actual distances considered in this study are rounded to nearest 100 km to be multiples of the span length

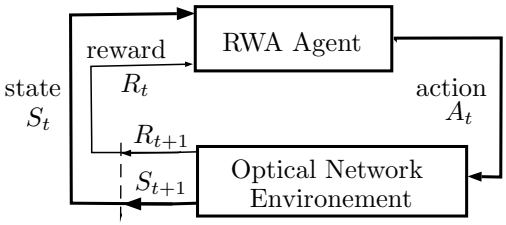

Fig. 1. A diagram of the MDP that defines the interactions between the RWA agent and the optical network

Events:
Arrival: *A service request arrives at time* $\mathsf{t_a}$ *with a source node* s*, destination node* d*, holding time* h *and bit rate* r*. Its either accepted or blocked. If accepted, its been allocated in a new or existing lightpath.*
Expiration: *If the expiring service (at time* $\mathsf{t_a}$ + h *) is the last remaining service in the lighpath, that lightpath is released.*

Constraints:
Routing constraint: *For each intermediate node* i *in the network, the flow conservation must be maintained.*
Wavelength continuity constraint: *If a lightpath is established between nodes* i *and* j *using wavelength* w*, the same wavelength* w *must be used on all the links along the path.*
Wavelength clash constraint: *A wavelength can be assigned to at most one lightpath on a given link, to avoid interference.*
Lightpath capacity constraint: *A service request can be added into a ligthpath if the available capacity* $\mathsf{C}_{available}$ *of the lightpath (the total capacity in Eq 1 - capacity allocated for any existing services in the lightpath) is greater than the service bit rate* r*.*

3. MULTI-OBJECTIVE REINFORCEMENT LEARNING FRAMEWORK FOR PARAMETER OPTIMIZATION

In RL, single objective optimization problems are commonly represented as a finite Markov decision process (MDP) [27]. This problem formulation is represented by the tuple $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$ with state-space $\mathcal{S}$, action space $\mathcal{A}$, scalar reward function $\mathcal{R}(S, A)$, and consists of an agent interacting with its environment at a series of time steps $1, 2, \ldots, t-1, t, t+1, \ldots$. The agent's goal is to learn an optimal policy Π^*, a functional mapping from the observed current state $S_t \in \mathcal{S}$ of the environment to the optimal action A^* $(\Pi^* : S \rightarrow A^*)$. A numerical reward $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ is provided for each action $A_t \in \mathcal{A}$, providing numerical feedback to the agent detailing how effective each action is. Fig. 1 depicts the MDP for an RL agent solving RWA in an optical network. The agent aims to maximize the cumulative future reward G_t for time-step t, defined as follows

$$G_t = \sum_{\tau=t+1}^{T} \kappa^{\tau-t-1} R_\tau\,, \quad \textbf{(2)}$$

where T denotes the total number of time-steps, and $\kappa \in [0, 1)$ denotes the discount factor [27].

In general, a multi-objective optimization problem (MOP) considers multiple objective functions simultaneously and the *Pareto-optimality* of the solutions is defined as the set of *non-dominated* solutions, where *dominance* relation is formulated as follows [28]

$$\mathsf{F}_i(x_1) \leq \mathsf{F}_i(x_2)\ \forall i \in \{1, \ldots, M\}\ \wedge$$
$$\exists j \in \{1, \ldots, q\} : \mathsf{F}_j(x_1) < \mathsf{F}_j(x_2)\,, \quad \textbf{(3)}$$

where x_1 and x_2 represent solutions in decision space and the respective values in objective space are represented by M objective functions $F_1, ..., F_i, ..., F_M$. The first condition states that the objective values of x_1 are no worse than those of x_2 in all objectives. The second condition describes that the objective values of solution x_1 are strictly better than at least one of those of solution x_2. If any of the two conditions are violated, the solution x_1 does not dominate the solution x_2. Otherwise, we can claim x_1 dominates x_2 [2]. The set of non-dominated solutions at the end of an optimization process form the *Pareto front*.

In an RL setting, a MOP is translated into a multi-objective Markov decision process (MOMDP) [18]. A MOMDP can be represented by $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$ parameters, similarly to a single-objective MDP as described above with the exception on $\mathcal{R} = [R_1, ..., R_i, ..., R_M]$ being a vector reward representing rewards for individual objectives $F_1, ..., F_i, ...F_M$, and with additional parameters. Namely, these are the space of preference vectors $\mathcal{Z}$ and the N preference functions $f_\omega(R)$, which each produces a scalar total reward $\mathcal{R}^{\mathrm{k}}_{\mathrm{tot}}$ for the respective preference $\omega^k = [\omega^k_1, \omega^k_2, ..., \omega^k_M] \in \mathcal{Z}$, $k \in \{1, \cdots, N\}$. Each such preference represents the relative importance of the M objectives as decided by the advisory (i.e. network operators) [29]. For the class of MOMDPs with linear preference functions, i.e., $f_\omega(R) = \omega | R$ (S, A) and ω^{k} is fixed to a constant value, this MOMDP will collapse into a standard MDP. On the other hand, if we consider all possible returns from a MOMDP for N preferences, we will have a Pareto front of rewards.

Let us consider a set of uniformly spread preference vectors $\omega^1, \cdots, \omega^k, \cdots, \omega^N$, for example, $[1, 0]$, $[0.9, 0.1]$, $\cdots$, $[0, 1]$ for a bi-objective problem, as shown in Fig 2. Thus, the original MOP is converted into N scalar optimization sub-problems by the weighted sum approach [20]. The objective function of the k^{th} sub-problem g_k is shown as follows [30]:

$$\min g(x \mid \omega^k) = \sum_{i=1}^{M} \omega^k_i F_i(x)\,. \tag{4}$$

Each of these subproblems g^k can then be represented by a standard MDP. In this work, the agent learns how to choose a set of optical network operational parameters given the state of the network on each timestep. The key components of our RL model are as follows.

Episode An episode consists of a series of timesteps. In each training episode we begin with an empty network and sequentially receive non-expiring requests at a rate of one request per timestep, which the agent aims to service.

State Representation In general, the state representation is a subset of the complete state of the optical network environment (consists of demand data, network utilization data, network graph features, optical network physical layer status such as noise to signal ratio (NSR) of the links etc.) depending on the specific optimization problem. For the RWA problem considered in this work, the state is described by the demand represented by the source and destination node ids and link utilization represented by the provisioned number of services.

Action The action A of RL corresponds to the vector of decision variables $X = [x_1, x_2, ..., x_n]$ in the considered optimization problem. Depending on the particular optimization problem within optical networks, the specific elements of the action vector varies. In this work, the action space is defined as the choice of a route out of the $k-$shortest paths[3] and a wavelength channel used to service a given demand.

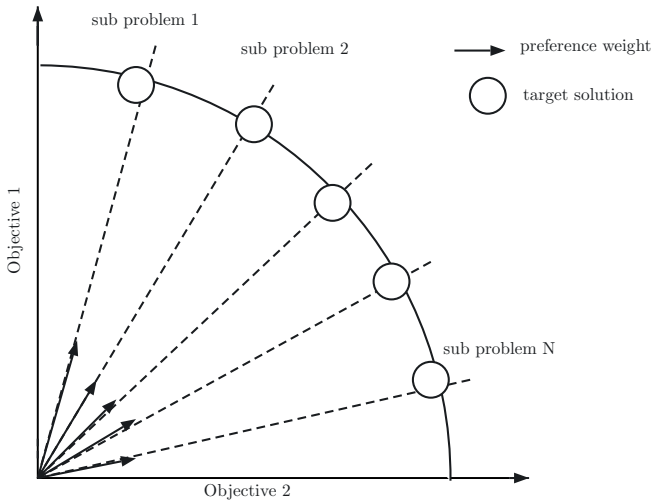

Fig. 2. An example of an approximated Pareto front for a bi objective optimization problem consisting the target solutions for the N sub problems

Reward Following the definition of MOMDP and Eq. (4), the total reward at a timestep t corresponding to subproblem k can be defined as follows, where $R_{t,i}(S_{t-1}, A_{t-1})$ is the individual reward for objective i at timestep t :

$$R^k_{\mathrm{tot}(t)}(S_{t-1}, A_{t-1}) = f_{\omega^k}(R_t) = \sum_{i=1}^{M} \omega^k_i R_{t,i}(S_{t-1}, A_{t-1})\,. \tag{5}$$

Algorithm 1. Multi-objective parameter optimization process

1) Initialize to constant values: the number of episodes K, number of steps in episode T, learning rate $\alpha \geq 0$, policy parameter $\theta \in R^d$, preference functions $f_{\omega^k}(R)$, $\omega^k = [\omega^k_1, \omega^k_2, ..., \omega^k_M]$ where $k = 1, \cdots, k, \cdots, N$, a differentiable parametric policy $\pi(A|S, \theta)$
for preference $f_{\omega^k}(R)$, $k = 1, 2, .., N$ **do**
  for episode $i = 1, 2, ..., K$ **do**
    for each step $t = 1, 2, ...T$ in episode **do**
      2) Generate action A_t for state S_t based on the policy $\pi(S_t|A_t, \theta)$
      3) Calculate $R^k_{\mathrm{tot}(t)}$ according to Eq. (5)
      4) Calculate G_t corresponding to $R^k_{\mathrm{tot}(t)}$ using Eq. (2)
      5) $\theta_{t+1} := \theta_t + \alpha G_t \nabla_\theta \ln \pi(A_t|S_t, \theta)$
6) Run the trained RL agents to generate a population of RWA solutions $P = [x_1, x_2, ..., x_n]$
7) Pareto front PF := Pareto sorting of solutions P based on Eq. (3)

Algorithm 1 outlines the steps in this approach. Firstly, initialization in Step 1) sets values to RL hyperparameters number

[2]This definition holds for a minimization problem, in the case of maximization problem the inequalities should be reversed.

[3]k-shortest paths between all the source and destination nodes are calculated using Dijkstra's algorithm [31].

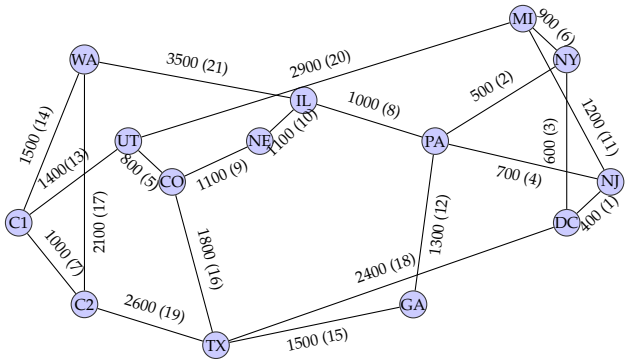

Fig. 3. Real-world backbone network topology NSFNET. The edge weights of the graph correspond to the distances given in km rounded to nearest 100 km.

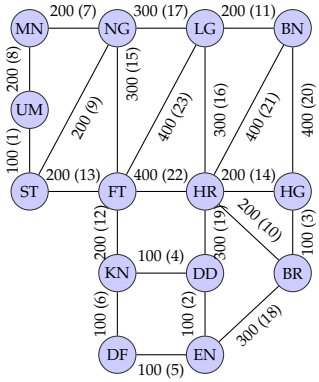

Fig. 4. Real world backbone network topology DTAG. The edge weights of the graph correspond to the distances given in km rounded to nearest 100 km.

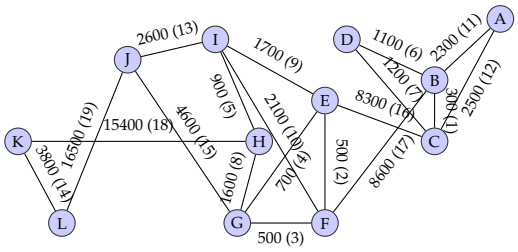

Fig. 5. Real world backbone network topology GB. The edge weights of the graph correspond to the distances given in km rounded to nearest 100 km.

of episodes K steps in one episode T, policy parameter $\theta \in R^d$ which consists of the weights of the neural network (NN) that presents RL policy network, learning rate α which balances between the fast convergence and overshooting for the NN, and the preference functions $f_{\omega^k}(R)$ for N preferences. RL hyperparameters can be chosen based on the values from the literature or through a parameter tuning process taken place prior to the algorithm run. Preference functions are defined to cover the entire range in the preference parameter space such that the set of sub-problems can collectively describe the original multi-objective problem (as shown in Fig. 2). Steps 2-5 describe the training process for the RL agents. Algorithm 1 loops through each preference function (the first For loop) for each episode (the second For loop) and each time step within an episode (the third For loop) by generating respective state (S_t) action (A_t) pairs and receiving respective reward vector ($\mathcal{R}_\sqcup = [R_{t(1)}, ..., R_{t(M)}]$) for the M objectives from the network simulator in Step 2). In Step 3) using this reward vector, the total scalar reward value ($R_{tot(t)}$) is calculated according to Eq. (5). Then the corresponding cumulative future rewards are calculated based on Eq. (2) in Step 4). Using the calculated cumulative reward value G_t, the vector of policy parameters θ_{t+1} is updated using a gradient update process [32] in Step 5). After the training process, in Step 6), the trained RL agents are run with unseen traffic to generate a population P of RWA solutions x_i. Then the Pareto sorting operation is performed for this population of solutions by removing the dominated solutions (Eq. (3) in Step 7), leaving the Pareto front of the solutions.

4. SIMULATION SET-UP

A. Network Topology

We consider three real-world core network topologies commonly referred in literature [1, 7, 12] for benchmarking, namely NSFNET depicted in Figure 3, DTAG in Figure 4, and GB in Figure 5.

B. Traffic Simulation

We consider a traffic model with non-expiring requests similar to Vincent et al. [26]. Bidirectional symmetric traffic is assumed and we consider the uniform-all-to-all model [1]. For a network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a set of nodes $\mathcal{V} \triangleq \{v_1, v_2, \ldots, v_N\}$ and a set of edges $\mathcal{E}$, source nodes $v_i \in \mathcal{V}$ and destination nodes $v_j \in \mathcal{V}$ $(i \neq j)$; the uniform traffic matrix $\widehat{T}_{\text{unif}}$ is defined as [1]

$$\widehat{T}_{\text{unif}}: \quad \forall \left\{v_i, v_j\right\} \in \mathcal{V}: \quad T_{ij} = \frac{1}{N(N-1)}, \tag{6}$$

where $N \triangleq |\mathcal{V}|$ is the total number of nodes in a given network.

C. Computational Set-Up

C.1. Network Simulator Set-up

The network simulation is developed on the open source Optical RL Gym library [33]. We use a fixed bit rate of 100 Gbps for each service request. Following the dynamic RWA model described in Section 2 B, lightpaths are modeled as having a given capacity (Eq. (1)), meaning that an existing lightpath can be used to service multiple requests between the same two nodes as long as there is sufficient spare capacity and other RWA constraints are obeyed. Accordingly, the Step 2 in Algorithm 1 can be elaborated. The network simulator decides whether to accept or reject a chosen action describing a route and a wavelength for a new lightpath or an existing one. If the chosen action represents a new lightpath request and satisfies the constraints of dynamic RWA problem (Section 2 B), a new lightpath is set up and the service is accepted. Similarly, in the case of the request is with respect to an existing lighpath, the service is accepted as long as the constraints described in Section 2 B are satisfied. Otherwise, the service request is blocked. We consider the holding time h to be extremely large such that the services are not expired

during the simulation. We adhere this model following from the sequential loading technique employed in the optical network literature [13, 26]. We consider $k = 5$ meaning that up to the $5^{\text{th}}-$shortest path can be chosen. Additional settings for bi-objective RWA and multi objective RWA are described in details in Section 5 and 6 respectively.

C.2. RL Agent Set-Up

RL agents are trained using an implementation of PPO provided by the Stable Baselines 3 library [34]. The hyperparameters considered in the training of the models for the three benchmark topologies are : discount factor $\kappa = 0.99$, learning rate of 1.57×10^{-5}, batch size equals to 16 and a network architecture of 2 layers of 128 neurons. 10 million timesteps in total are considered corresponding to 5000 episodes. Each episode consists of 2000 timesteps following the computationally efficient scaling approach proposed in Nevin et al. [12]. The scaled down episode sizes during the training process is observed to be effective for a larger episode sizes in running the RL agents. Hence, the computational cost of the training process can be considered as the cost to run 10 million timespteps multiplied by the number of preference functions considered in the specific multi objective optimization problem. For validation, 30 episodes with the episode size of 10000 timesteps are considered. 10000 timesteps are chosen as the upper limit for the serviceable requests based on the simulation results for the benchmark topologies. All other parameters are equal to the defaults in Stable Baselines 3.

5. SIMULATION RESULTS FOR RWA WITH OBJECTIVES ACCEPTED SERVICES AND TRANSMITTERS

Section 3 described the action space and state representation for RWA problem in general. In this section, we consider a specific reward function to solve the RWA problem with respect to the bi-objective version of maximizing the number of accepted services and minimizing the number of transmitters. These objectives are motivated by the performance and cost goals of the network. In a network, controlling the number of transmitters can be a viable method for reducing financial costs. Moreover, this strategy remains relevant in situations where the budget is flexible, as it enables the trade-off between the performance gains and the additional costs incurred.

Reward Based on the weighted sum approach discussed in Section 3, the original problem is converted into scalar sub problems (Eq. (4)). Accordingly, the respective preference vectors corresponding to the reward functions for each sub-problem (Eq. (5)) are chosen to be spread linearly across the preference space Z . The preference vectors considering the objectives of throughput for the m light-paths p_j (Eq. (7)) and the number of transmitters across l nodes N_i (Eq. (8)) are $[0.1, 0.9]$, $[0.2, 0.8]$, $[0.3, 0.7]$, $[0.4, 0.6]$, $[0.5, 0.5]$, $[0.6, 0.4]$, $[0.7, 0.3]$, $[0.8, 0.2]$, $[0.9, 0.1]$. Objective throughput maximization is further simplified as the maximization of the number of accepted services as all the services in the considered simulation correspond to a fixed bit rate of 100 Gbps.

$$\begin{aligned} \text{Find:} \quad & \boldsymbol{X} = [x_1, x_2, \ldots, x_n] \\ \text{Maximize:} \quad & F_{\mathsf{T}}(\boldsymbol{X}) = \sum_{j=0}^{m} \mathsf{T}_{p_j}, \end{aligned} \tag{7}$$

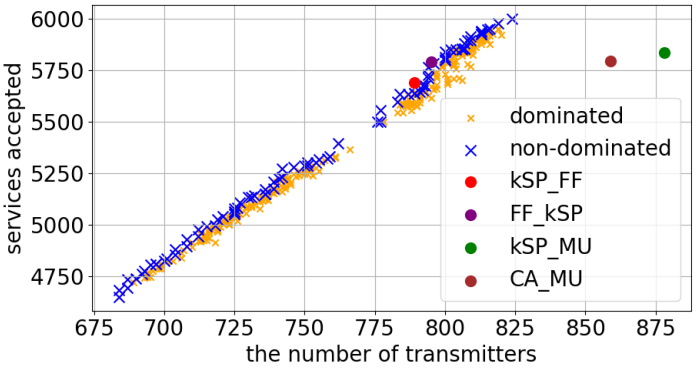

Fig. 6. Approximated Pareto front (consisting non-dominated points in blue) and dominated points (orange) at the end of the optimization process for bi objective RWA for NSFNET topology. Additionally, the results for RWA heuristics kSP-FF, FF-kSP, kSP-MU and CA-MU are presented for benchmarking purposes.

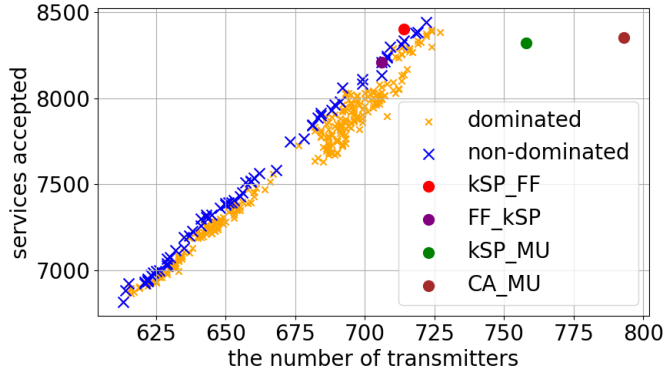

Fig. 7. Approximated Pareto front (consisting non-dominated points in blue) and dominated points (orange) at the end of the optimization process for bi objective RWA for DTAG topology. Additionally, the results for RWA heuristics kSP-FF, FF-kSP, kSP-MU and CA-MU are presented for benchmarking purposes.

$$\begin{aligned} \text{Find:} \quad & \boldsymbol{X} = [x_1, x_2, \ldots, x_n] \\ \text{Minimize:} \quad & F_{\text{Transmitters}}(\boldsymbol{X}) = \sum_{i=0}^{l} N_i \,. \end{aligned} \tag{8}$$

Figures 6, 7 and, 8 present the results of objectives representing the minimization of the number of transmitters and maximization of accepted services for NSFNET, DTAG and GB topologies respectively. Results for all three benchmark topologies show a strong positive correlation of the number of transmitters and services accepted within the considered region. Pearson's correlation coefficient of 0.997 is observed for NSFNET while DTAG and GB correspond to values of 0.998 and 0.989. Since our objectives are to maximize the number of services accepted (Eq. (7)) and minimize the number of transmitters (Eq. (8)) this reads as a negative correlation. On a network with a given load, the financial cost can be reduced by limiting the number of transmitters up to a number which corresponds to a specified number services accepted, fulfilling the respective load requirement. Furthermore, when the budget is flexible this knowledge can be still useful to trade-off the performance gain against additional incurred cost. On average, an addition of a transmitter

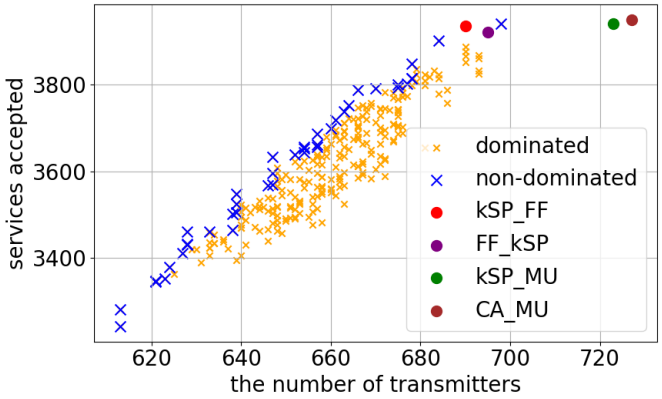

Fig. 8. Approximated Pareto front (consisting non-dominated points in blue) and dominated points (orange) at the end of the optimization process for bi objective RWA for GB topology. Additionally, the results for RWA heuristics kSP-FF, FF-kSP, kSP-MU and CA-MU are presented for benchmarking purposes.

corresponds to a gain of 10, 17.5, and 7.5 accepted services for NSFNET, DTAG and GB, respectively. Moreover, we benchmark the performance of the RL solutions against four state-of-the-art RWA heuristics: $k-$shortest path first fit (kSP-FF), first fit $k-$shortest path (FF-kSP), k-shortest path most-used (kSP-MU) and congestion aware most used (CA-MU) [26]. [4] For the three benchmark networks FF-kSP lies on the Pareto front and kSP-FF lies for NSFNET and DTAG while kSP-MU and CA-MU are being dominated. Results suggest that RL has found the optimal solutions in the same range as the state-of-the-art heuristics while in addition being able to provide a range of optimal solutions instead of a single solution.

6. SIMULATION RESULTS FOR RWA OPTIMIZATION FOR ACCEPTED SERVICES, AVAILABILITY AND TRANSMITTERS

In this section, we present results for the simultaneous maximization of the number of accepted services, the availability and the minimization of the number of transmitters. Based on the literature, network availability can be calculated using the failures of all the elements in the network, meaning all fiber links, transmitters, receivers and amplifiers etc. Availability is conventionally modeled using a Markov availability model [35] where the availability is represented by available and failure states and the time to repair parameter. We consider a more simplistic model referred in Archi et al. [36] where link availability is modeled with a fixed failure rate β which is time independent and nodes are modeled as utterly reliable components (i.e failure probability of a node is zero). According to this model, the event of path failure is resultant entirely from a failure of any of the links the path traverses. Hence, the probability of path availability can be obtained by the probability of the intersection of the events corresponding to the availability of the links through which the path traverses. Accordingly, the link availability L_j and path availability A_{p_i} for this metric can be described by Eq. (9) and Eq. (10) [37], respectively, with β being the failure probability per 1 km and l the number of links the path p_i traverses. It yields:

$$\text{Link availability:} \quad L_j \triangleq 1 - \left(\beta \times \text{link length}_j\right) \tag{9}$$

$$\text{Path availability:} \quad A_{p_i} \triangleq \bigcap_{j=1}^{l} L_j \,. \tag{10}$$

Subsequently, we can define the objective of maximizing availability score A for the network taken as the average across the availability scores for all the occupied light-paths as follows:

$$\begin{aligned} \text{Find:} \quad & \boldsymbol{X} = [x_1, x_2, \ldots, x_n] \\ \text{Maximize:} \quad & F_{\mathrm{A}}(\boldsymbol{X}) = \frac{1}{m}\sum_{i=0}^{m} \mathrm{A}_{p_i}\,, \end{aligned} \tag{11}$$

where m denotes the total number of occupied light-paths in the network Eq. (1).

The state and action spaces for this study are the same as the work presented in Section 5, while, the reward function is extended to consider the additional objective of network availability.

Reward Similar to the approach in Section 5 the preference vectors are sampled to be spread across the preference space $\mathcal{Z}$. The values for preference vectors, considering the objectives of maximization of accepted services (Eq. (7)), maximization of availability (Eq. (11)) and minimization of the number of transmitters (Eq. (8)), are $[0.1, 0.1, 0.8]$, $[0.1, 0.8, 0.1]$, $[0.8, 0.1, 0.1]$, $[0.2, 0.2, 0.6]$, $[0.2, 0.6, 0.2]$, $[0.6, 0.2, 0.2]$, $[0.45, 0.45, 0.1]$, $[0.45, 0.1, 0.45]$, $[0.1, 0.45, 0.45]$, $[0.33, 0.33, 0.33]$, respectively. The unit failure probability (per 1 km) for the fiber links β is set to 10^{-7}.

We have performed constrained optimization with a lower bound on accepted services and objectives representing the minimization of the number of transmitters and maximization of availability for NSFNET, DTAG and GB topologies[5]. Specifically, the constraints applied are a minimum number of accepted services of 5000, 7000 and 3500 for the NSFNET, DTAG and GB topologies, respectively. The respective rounded values are chosen heuristically to consider the range of choices lie between the states of a network that operates on 80% to 100% of its total capacity. Depending on the specific throughput/number of accepted services requirement for the considered optical network this constraint can be modified to contain a different range of options (for example, for a often under-loaded network operates on 50% of its total capacity, a smaller throughput constraint will be more suitable as it will provide choices to downsize the network to 50% while reducing the costs and increasing the availability, etc.).

These results can be useful in the case where network operators need to decide the use of additional transmitters to increase availability and throughput (i.e., accepted services=5942, availability = 99.9969, transmitters = 816 in Fig. 9) or considering downsizing of an under-loaded network down to a certain throughput level to reduce the number of transmitters while

[4] kSP-FF searches for a lightpath that can support the current request, starting with the shortest path and searching each channel sequentially until a valid lightpath is found. If a lightpath is not found for the shortest channel, the second-shortest path is searched and so on. In contrast, FF-kSP starts with the first channel slot and searches each of the shortest paths in order of length to find a lightpath that can support the current request. kSP-MU searches each channel in order of length, allocating the request to the most-used wavelength in the network at the current time. Similarly, CA-MU extends kSP routing by searching routes in an order determined not only by path length but network congestion [26].

[5] We include Figure 9 for NSFNET in the manuscript and due to space limitations we only present statistical results for DTAG and GB here.

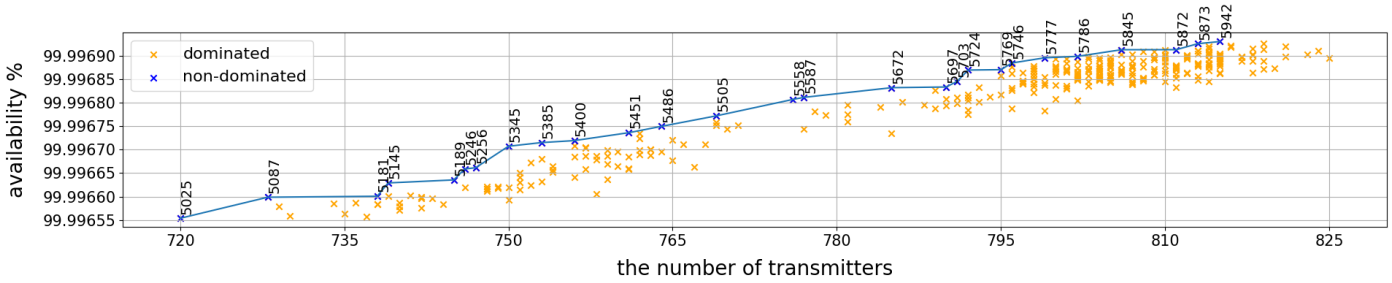

Fig. 9. Approximated Pareto front (consisting the non-dominated points in blue) and the dominated points (orange) at the end of the optimization process for NSFNET topology for objectives considering availability Eq. (11) and the number of transmitters Eq. (8) and a constraint of number of accepted services > 5000. The respective number of services accepted are annotated above the scatter points.

increasing availability (i.e., accepted services = 5087, availability = 99.9966, transmitters = 720 in Fig. 9). For the considered topologies, a strong positive correlation is observed between the number of transmitters and availability (hence, a strong negative correlation between the minimization of transmitters). Pearson's correlation coefficient of 0.991 is reported for NSFNET, while for DTAG and GB values of 0.989 and 0.871 are reported respectively. These are quite similar to the correlation reported in previous section between the transmitters and the services accepted. This further implies a strong positive correlation between the objectives of maximization of the number of accepted services and availability. In next section, we will investigate this further to observe any similarities between the learned policies.

7. INTERPRETATION OF THE LEARNED RL POLICY

It is important to consider how the learned policy can be interpreted, to improve operator confidence in the proposed solutions. Thus, we performed further simulations motivated by understanding what the RL agent has learned. Different policies could represent the preference criteria determined by the advisory (i.e, network operators, routing policy designers, etc.) in the network design and operation processes for different networks and times/modes of operation. For example, on a network receiving a large load requiring to operate on the full capacity, operators might give priority to operation on the maximum network capacity over reducing the cost while on a network which often is under-loaded it might be of higher priority to reduce the cost (i.e. minimize the number of transmitters) than operating on the maximum capacity. On the other hand, on a network which is built on a infrastructure which is severely prone to link failures, it would be of the highest importance to maintain the availability. To reflect these different scenarios, we consider following 3 preferences as depicted in Table 1 representing three different choices the advisory can make considering the priority to one objective over the other two. Preference A ($\omega^A = [\omega_1^A = 0.1,\ \omega_2^A = 0.1,\ \omega_3^A = 0.8]$) has higher priority for objective of minimization of the number of transmitters ($\omega_3 = 0.8$) and lower priorities to maximization of accepted services ($\omega_1 = 0.1$) and maximization of availability ($\omega_2 = 0.1$). Similarly, preference B and C corresponds to higher priority to objective maximization of availability and maximization of the accepted services respectively (Table 1).

Fig. 10 depicts the serviced requests distribution over the links for the NSFNET topology. The top sub-figure corresponds to preference A which has a preference weight of 0.1 for the objective of maximization of the number of services Eq. (7), 0.1 for the objective of maximization of availability Eq. (11) and 0.8 for the objective of minimization of the number of transmitters Eq. (8). The higher preference weight for minimization of transmitters resulted in considerably fewer services provisioned using fewer transmitters with a per link average of 580.0 and standard deviation of 45.3. Still, it did not reach 0 due to the non-zero preference weight for the other objectives Eq. (7) and Eq. (11).

Respectively, the middle sub-figure of Fig. 10 has preference weights of 0.1, 0.8, and 0.1 for the number of services, availability and transmitters respectively, representing preference B. As shown in Table 2, both these figures have significantly higher number of services across the links (with averages of 610.5, 614.7 and standard deviations of 85.6, 85.2 for the middle and bottom sub-figures, respectively) compared to the top sub-figure while the bottom sub-figure has the highest. This further suggests the similarities of the two objectives service maximization Eq. (7) and availability maximization Eq. (11) observed in Section 6. Moreover, the maximum (link 3) and minimum (link 16) utilized links are observed to be the same for the two cases. The availabil-

Table 1. Preferences A, B and C

id/preference	accepted services ω_1	availability ω_2	transmitters ω_3
preference A	$\omega_1^A = 0.1$	$\omega_2^A = 0.1$	$\omega_3^A = 0.8$
preference B	$\omega_1^B = 0.1$	$\omega_2^B = 0.8$	$\omega_3^B = 0.1$
preference C	$\omega_1^C = 0.8$	$\omega_2^C = 0.1$	$\omega_3^C = 0.1$

Table 2. Statistics for the learned policies for NSFNET for preferences A, B, and C

statistic/preference	A	B	C
per link services			
mean	580.0	610.5	614.7
SD	45.3	85.6	85.2
max	662.1 (link 1)	714.7 (link 3)	723.4 (link 3)
min	467.3 (link 16)	396.6 (link 16)	402.4 (link 16)
per wavelength services			
mean	121.8	128.2	129.1
SD	13.0	16.5	15.8
max	150.2 (WL 99)	157.4 (WL 65)	157.4 (WL 43)
min	88.4 (WL 92)	89.5 (WL 14)	94.1 (WL 27)

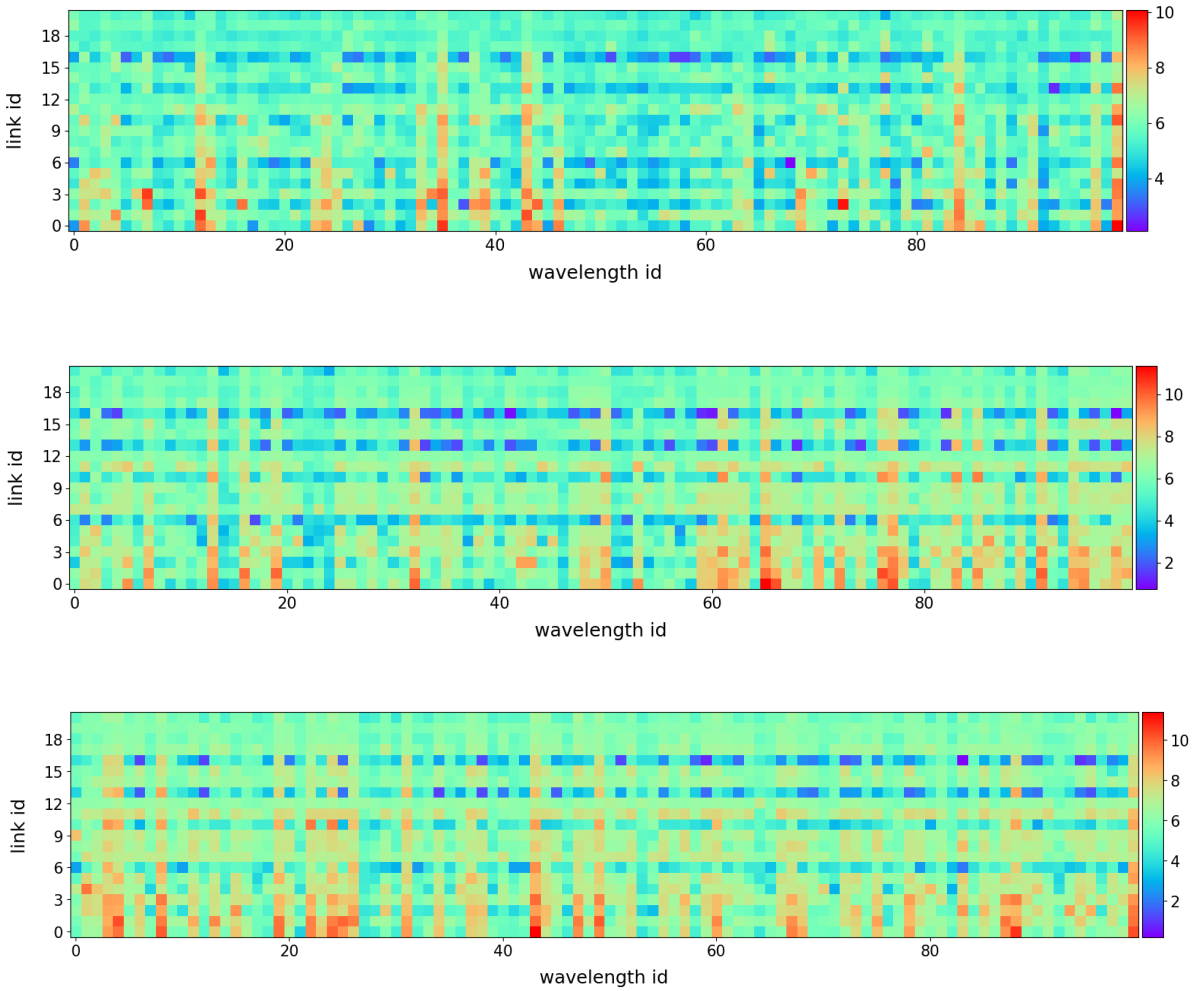

Fig. 10. Service distribution across links in for different preference functions for NSFNET depicting services accepted, availability and transmitters with preference A: $\omega^A = [\omega_1^A = 0.1,\ \omega_2^A = 0.1,\ \omega_3^A = 0.8]$(top), preference B: $\omega^B = [\omega_1^B = 0.1,\ \omega_2^B = 0.8,\ \omega_3^B = 0.1]$ (middle), and preference C: $\omega^C = [\omega_1^C = 0.8,\ \omega_2^C = 0.1,\ \omega_3^C = 0.1]$ (bottom). Link ids are as described in the NSFNET topology in Figure 3 and the color-bar describes the number of accepted services.

ity metric Eq. (11) considers the length of the fiber links as the basis for failure probability calculation, while the path capacity is also a function of the lengths of the links, as the link length affects SNR Eq. (1). Similar patterns are observed for DTAG and GB networks. For DTAG, link 19 and link 17 were the maximum and minimum utilized for both preference *B* and *C*. For GB, only the minimum utilized link (link 3) is shared between these preferences. Moreover, for both topologies, the average number of accepted services per link were similar for preference *B* and *C* (948.0 and 945.6 respectively for DTAG and 375.6 and 379.8 respectively for GB).

Moreover, Figure 11 presents the service distribution across the links for NSFNET for the three RL policies A, B and C comparative to k-shortest path (kSP) and congestion aware (CA) heuristic routing policies [26]. Strong similarities were observed between routing policies with respect to preference functions B and C (RL_B and RL_C) and with both heuristic policies. This is further evident by Pearson's correlation coefficient values being closer to 1 for RL_B, RL_C and kSP policy (0.89 for RL_B and kSP and 0.88 for RL_C and kSP) while the coefficients for RL and CA routing policy is slightly higher (0.94 for RL_B and CA and 0.95 for RL_C and CA). By maximizing the number services (throughput) and network availability, RL seems to learn to route being aware on the congestion levels as well as finding shortest paths. Moreover, RL_C also has shown positive correlations with heuristic routing policies (0.74 and 0.79 correlation with kSP and CA respectively). However, this is a weaker correlation than for the other two RL policies.

Similarly, Figure 12 presents the service distribution across the wavelengths for NSFNET for the three RL policies A, B and C comparative to the heuristic policies first-fit (FF) and most used (MU) [26]. There were no observable similarities between the complex RL policies and simple heuristic policies in terms of wavelength assignment. This was further evident by values close to 0 for both Pearson's correlation coefficient and Spearman's correlation coefficient. [6]

[6]Due to space limitations only the results for NSFNET is included. Similar results were observed for the other networks.

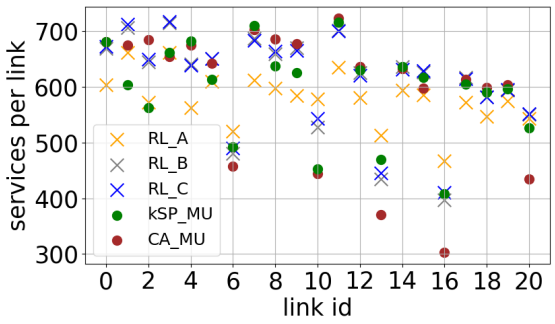

Fig. 11. Services distribution across links for NSFNET topology for A, B and C RL policies and shortest path (kSP) and congestion aware (CA) heuristic routing policies.

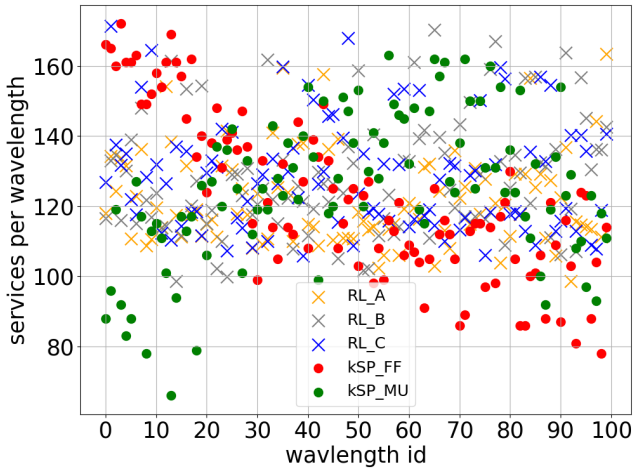

Fig. 12. Services distribution across wavelengths for NSFNET topology for A,B and C RL policies and first fit (FF) and most used (MU) heuristic wavelength policies.

8. GENERALIZATION TO NON-UNIFORM TRAFFIC

In order to investigate the generalizability of this RL model, we evaluate the RL agent trained on uniform traffic matrix (Eq. (6)) on a realistic non-uniform population based traffic distribution not seen during training. Let S and D be the discrete random variables denoting the source and the destination, respectively. Possible values each can take are $1, 2, \ldots, i, \ldots, k$ with k being the number of nodes. If r_i is the number of residents for the $i^{\text{th}}-$node, then we assume the probability of selecting the source is the population of the source as a fraction of the total population. The population based traffic matrix $T_{ij} = T_{ji}$ [12] is defined as follows:

$$T_{ij} = \frac{1}{2}\left(\frac{r_i}{\sum_k r_k}\frac{r_j}{\sum_{k\neq i} r_k} + \frac{r_j}{\sum_k r_k}\frac{r_i}{\sum_{k\neq j} r_k}\right). \quad (12)$$

As shown in Figure 13, the RL agent is able to learn a generalizable policy from the uniform traffic distribution during training, allowing it to perform well for a different, non-uniform traffic distribution without retraining. Therefore, the RL agent affords the operator a flexibility advantage over the heuristics, which need to be hand-tuned for each problem. This flexibility is one of the major advantages of RL-driven solutions.

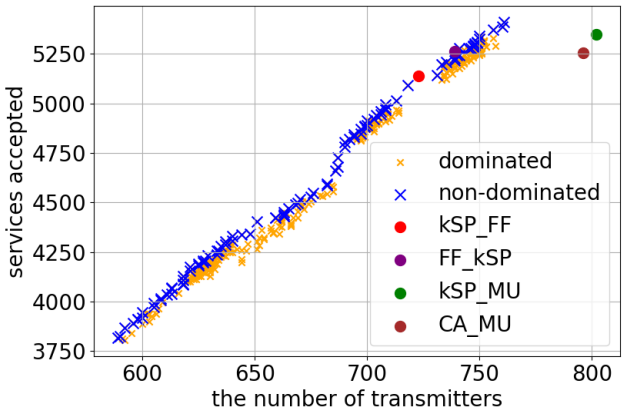

Fig. 13. Approximated Pareto front (consisting non-dominated points in blue) and dominated points (orange) for bi objective RWA for NSFNET topology for population based traffic (Eq. (12)). Additionally, the results for RWA heuristics kSP-FF, FF-kSP, kSP-MU and CA-MU are presented for benchmarking purposes.

9. CONCLUSIONS

A generalized reinforcement learning framework for multi-objective parameter optimization in optical networks is proposed. The proposed framework is applied to two multi-objective variants of the classical routing and wavelength assignment problem and its efficacy is demonstrated using a simulated network environment. Pareto frontiers of optimal solutions with respect to three objectives, namely the maximization of the number of accepted services, minimization of the number of transmitters, and the maximization of availability are constructed for three benchmark network topologies of varying scales, from country to continental to global-scale networks. These results provide insights for network designers and operators in decision making on network design, configuration and operation. Specifically, by providing network operators with a set of optimal solutions generated by multi objective optimization, defined by the Pareto front, operators can pick a solution based on their current priority. Benchmark results over the state-of-the-art RWA heuristics further suggest the ability of RL in finding the optimal solutions in the same range that of the heuristics with the significant added value owing the multiple optimal values. Computational cost in machine learning model training is significantly reduced using an effective scaling approach to reduce the RL episode size required for simulation during training. Additional investigation to interpret the policies learned by the RL agents suggests the similarities between the services accepted and the availability objectives and the dependency of these objectives on topological structure. This motivates further research into understanding the inter-dependencies among network parameters and performance metrics. Moreover, the generalizability of the learned policies to unseen non-uniform traffic is shown.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council Programme Grant TRANSNET [grant number EP/R035342/1].

OPEN ACCESS AND DATA AVAILABILITY STATEMENT

 Data underlying the results is available at https://doi.org/10.17863/CAM.91675.

REFERENCES

1. D. J. Ives, P. Bayvel, and S. J. Savory, "Routing, modulation, spectrum and launch power assignment to maximize the traffic throughput of a nonlinear optical mesh network," Photonic Netw. Commun. **29**, 244–256 (2015).
2. I. Roberts, J. M. Kahn, J. Harley, and D. W. Boertjes, "Channel power optimization of WDM systems following Gaussian noise nonlinearity model in presence of stimulated Raman scattering," J. Light. Technol. **35**, 5237–5249 (2017).
3. E. Virgillito, R. Sadeghi, A. Ferrari, A. Napoli, B. Correia, and V. Curri, "Network Performance Assessment with Uniform and Non-Uniform Nodes Distribution in C+L Upgrades vs. Fiber Doubling SDM Solutions," in *2020 International Conference on Optical Network Design and Modeling (ONDM),* (2020), pp. 1–6.
4. P. Poggiolini, G. Bosco, A. Carena, R. Cigliutti, V. Curri, F. Forghieri, R. Pastorelli, and S. Piciaccia, "The LOGON strategy for low-complexity control plane implementation in new-generation flexible networks," in *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC),* (2013).
5. C. M. D. Uzunidis, E. Kosmatos and A. Stavdas, "Strategies for upgrading an operator's backbone network beyond the C-band: Towards multi-band optical networks," IEEE J. Photonics (2021).
6. P. Bayvel, R. Maher, T. Xu, G. Liga, N. A. Shevchenko, D. Lavery, A. Alvarado, and R. I. Killey, "Maximizing the optical network capacity," Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci. **374**, 20140440 (2016).
7. X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. Yoo, "DeepRMSA: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," J. Light. Technol. **37**, 4155–4163 (2019).
8. R. Weixer, S. Kuhl, M. Manuel, B. Spinnler, W. Schairer, B. Sommerkorn-Krombholz, and S. Pachnicke, "A Reinforcement Learning Framework for Parameter Optimization in Elastic Optical Networks," in *Eur. Conf. Opt. Commun. (ECOC),* (2020).
9. P. Bayvel, R. Luo, R. Matzner, D. Semrau, and G. Zervas, "Intelligent design of optical networks: which topology features help maximise throughput in the nonlinear regime?" in *European Conference on Optical Communications (ECOC),* (2020).
10. S. Nallaperuma, N. A. Shevchenko, and S. J. Savory, "Parameter optimisation for ultra-wideband optical networks in the presence of stimulated Raman scattering effect," in *2021 International Conference on Optical Network Design and Modeling (ONDM),* (2021), pp. 1–6.
11. N. A. Shevchenko, S. Nallaperuma, and S. J. Savory, "Maximizing the information throughput of ultra-wideband fiber-optic communication systems," Opt. Express **30**, 19320–19331 (2022).
12. J. W. Nevin, S. Nallaperuma, N. A. Shevchenko, Z. Shabka, G. Zervas, and S. J. Savory, "Techniques for applying reinforcement learning to routing and wavelength assignment problems in optical fiber communication networks," J. Opt. Commun. Netw. **14**, 733–748 (2022).
13. N. D. Cicco, E. F. Mercan, O. Karandin, O. Ayoub, S. Troia, F. Musumeci, and M. Tornatore, "On deep reinforcement learning for static routing and wavelength assigment," IEEE J. Sel. Top. Quantum Electron. (2022).
14. A. Samadian, X. Wang, M. Razo, A. Fumagalli, and C. Lee, "Two conflicting optimization problems in wdm networks: Minimizing spectrum fragmentation and maximizing quality of transmission," in *2018 IEEE International Conference on Communications (ICC),* (2018).
15. I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical WAN's," IEEE Trans. Commun. **40**, 1171–1182 (1992).
16. B. Jaumard, C. Meyer, and B. Thiongane, "Comparison of ILP formulations for the RWA problem," Opt. Switch. Netw. **4**, 157–172 (2007).
17. H. Zang and J. P. Jue, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," Opt. Networks Mag. **1**, 47–60 (2000).
18. B. E. e. a. Hayes C.F., Rădulescu R., "A practical guide to multi-objective reinforcement learning and planning." Auton Agent Multi-Agent Syst **36** (2021=2).
19. C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," IEEE Transactions on Syst. Man, Cybern. Syst. **45**, 385–398 (2015).
20. K. Li, T. Zhang, and R. Wang, "Deep reinforcement learning for multi-objective optimization," IEEE Trans. Cybern. (2020).
21. H. Kaur and M. Rattan, "Improved offline multi-objective routing and wavelength assignment in optical networks," Front. Optoelectronics **12**, 433–444 (2019).
22. Z. Xu, Q. Xu, J. Lv, T. Ma, and T. Chen, "An adaptive multiobjective genetic algorithm with multi-strategy fusion for resource allocation in elastic multi-core fiber networks," Appl. Sci. **12** (2022).
23. S. Nallaperuma, N. A. Shevchenko, and S. J. Savory, "A pareto-optimality based multi-objective optimisation approach to assist optical network (re-)design choices," in *2021 European Conference on Optical Communication (ECOC),* (2021), pp. 1–4.
24. O. Şeker, M. Bodur, and H. Pouya, "Routing and wavelength assignment with protection: A quadratic unconstrained binary optimization approach," (2020).
25. P. Poggiolini, "The GN model of non-linear propagation in uncompensated coherent optical systems," J. Light. Technol. **30**, 3857–3879 (2012).
26. R. J. Vincent, D. J. Ives, and S. J. Savory, "Scalable capacity estimation for nonlinear elastic all-optical core networks," J. Light. Technol. **37**, 5380–5391 (2019).
27. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
28. S. Nallaperuma, N. A. Shevchenko, and S. J. Savory, "A Pareto-optimality based multi-objective optimisation approach to assist optical network (re-)design choices," in *European Conference on Optical Communication, ECOC 2021, Bordeaux, France, September 13-16, 2021,* (IEEE, 2021), pp. 1–4.
29. R. Yang, X. Sun, and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS),* (Curran Associates Inc., Red Hook, NY, USA, 2019).
30. Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," IEEE Transactions on Evol. Comput. **11**, 712–731 (2007).
31. E. W. Dijkstra, "A on two problems in connexion with graphs," Numer. mathematik **1**, 269–271 (1959).
32. R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems,* vol. 12 S. Solla, T. Leen, and K. Müller, eds. (MIT Press, 1999).
33. C. Natalino and P. Monti, "The Optical RL-Gym: An open-source toolkit for applying reinforcement learning in optical networks," in *Int. Conf. Transparent Opt. Netw. (ICTON),* (2020), p. Mo.C1.1.
34. A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," J. Mach. Learn. Res. **22**, 1–8 (2021).
35. L. Wosinska, D. Colle, P. Demeester, K. Katrinis, M. Lackovic, O. Lapcevic, I. Lievens, G. Markidis, B. Mikac, M. Pickavet, B. Puype, N. Skorin-Kapov, D. Staessens, and A. Tzanakaki, *Network Resilience in Future Optical Networks* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009), pp. 253–284.
36. D. Arci, G. Maier, A. Pattavina, D. Petecchi, and M. Tornatore, "Availability models for protection techniques in WDM networks," in *Fourth International Workshop on Design of Reliable Communication Networks, 2003. (DRCN 2003). Proceedings.*, (2003), pp. 158–166.
37. J. Segovia, E. Calle, P. Vila, J. Marzo, and J. Tapolcai, "Topology-focused availability analysis of basic protection schemes in optical transport networks," J. Opt. Netw. **7**, 351–364 (2008).