Parameterized reinforcement learning for optical system optimization

OPEN ACCESS
IOP Publishing
Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **54** (2021) 305104 (11pp)
https://doi.org/10.1088/1361-6463/abfddb

Parameterized reinforcement learning for optical system optimization

Heribert Wankerl[1,2,*], Maike L Stern[1], Ali Mahdavi[1], Christoph Eichler[1] and Elmar W Lang[2]

[1] OSRAM Opto Semiconductors, Regensburg, Germany
[2] University of Regensburg, Regensburg, Germany

E-mail: heribert.wankerl@osram-os.com

Received 21 January 2021, revised 12 April 2021
Accepted for publication 4 May 2021
Published 18 May 2021

Abstract

Engineering a physical system to feature designated characteristics states an inverse design problem, which is often determined by several discrete and continuous parameters. If such a system must feature a particular behavior, the mentioned combination of both, discrete and continuous, parameters results in a challenging optimization problem that requires an extensive search for an optimal system design. However, if the corresponding inverse design problem can be reformulated as a parameterized Markov decision process, reinforcement learning (RL) provides a heuristic framework to solve it. In this work, we use multi-layer thin films as an example of the aforementioned optimization problems and consider three design parameters: Each of the thin film layer's dielectric material (discrete) and thickness (continuous), as well as the total number of layers (discrete). While recent methods merely determine the optimal thicknesses and—less commonly—the layers' materials, our approach optimizes the total number of stacked layers as well. In summary, we further develop a Q-learning variant to solve inverse design optimization and thereby outperform human experts and current approaches like needle-point optimization or naive RL. For this purpose, we propose an exponentially transformed reward signal that eases policy search and enables constrained optimization. Moreover, the learned Q-values contain information about the optical properties of multi-layer thin films, which allows us a physical interpretation or what-if analysis and thus enables explainability.

Keywords: machine learning, reinforcement learning, inverse design problem, optics, multi-layer thin-film, optimization

(Some figures may appear in colour only in the online journal)

1. Introduction

In many fields of physics and engineering, the design of a system is determined by (design) parameters. In recent years, the numerical prediction of the physical behavior of given designs by forward simulations has become faster due to advances in computational sciences. Conversely, the search for optimal design parameters of a system that features a required behavior remains a challenging inverse design problem. Actually, it becomes even more ambitious as breakthroughs in fabrication methods and material sciences increased the number of accessible design parameters as well. Notably, a linear increase of tunable parameters results in an exponential increase of the search space volume that needs to be explored—an effect that is often referred to as the *curse of dimensionality* [5]. By

* Author to whom any correspondence should be addressed.

solving inverse design problems regarding optical nanostructures [36, 37, 40, 65, 70], meta-surfaces [27, 38, 69], integrated photonics [18, 58], and thin films [36], deep learning [19, 33] has proven to potentially master the curse of dimensionality.

Basically, deep neural networks (DNNs) can be thought of as approximators for arbitrary nonlinear functions with any desired accuracy [12, 13]—in natural sciences this is often exploited to obtain differentiable surrogate models of physical processes. As such, they can be trained to estimate the physical behavior given a design but not vice-versa due to the ambiguous relations between a target behavior and the possibility of various corresponding design solutions. Therefore, DNNs are used to surrogate the forward simulations of given designs in order to implement gradient-based [46], generative [37] or evolutionary optimization approaches [23]. The latter aim to compensate for prediction errors of DNNs [17, 52], which are used for regression tasks. In general, the sufficient training of DNNs in a supervised manner requires extensive hyperparameter optimization and costly pre-selection of an extensive dataset to properly reflect the design space [23, 33].

Instead of being directly trained to approximate the underlying physical problem, DNNs can be implemented to estimate the expected future reward of a particular design. Here, the reward of a design is determined by its performance of approximating a particular target behavior. Reinforcement learning (RL) utilizes such so-called value function approximations to enforce high rewards by adapting design parameters. This sequential adaption of parameters corresponds to traversing a trajectory of designs in the search space. Importantly, the RL agent can learn an optimal policy from delayed rewards and is thus able to globally optimize non-convex functions. The corresponding formalism, notations and technical terms are explicitly introduced in sections 3 and 4. Although known since the 1950s, RL recently gained attention for solving inverse design problems when it beats human skills at games like chess [54] and Go [51, 53]. Since then, RL has been used in optics and nanophotonics to deduce compact integrated photonic devices [3] like on-chip silicon beam splitters (T-junctions) [2], based on binary matrix representations of meta-materials. Moreover, researchers reported Q-learning [63], a variant of RL, to stabilize the phases in coherent beam combining applications [61] and the operation of mode-locked lasers [56] by means of influencing operating conditions like driving voltage, wave plates or polarizers, respectively. The same method was implemented to optimize the color generation from dielectric nanostructures [25, 48] as well as to find the optimal thickness configuration of multi-layer thin films [29]. In general, most of the recently proposed and almost all of the previously mentioned RL approaches that solve inverse design problems implement variants of Q-learning so as to find the optimal design parameters. By construction, Q-learning requires each possible design parameter to be discrete, although they may appear to be continuous in nature. In such cases, a vectorization and discretization of continuous parameters is inevitable. This often not only requires many dimensions and prohibits the integral solving of specific tasks, it also implies an unphysical degradation of the underlying problem. To overcome the mentioned inadequacies, we propose to formulate design spaces that accommodate both, continuous and discrete, parameters as so-called parameterized actions spaces. A brief summary of the corresponding research history is provided in section 2.

In this work, the integral optimization of multi-layer thin films is taken as an example to demonstrate how parameterized RL can lead to very intuitive implementations of physical problems that feature both, inter-dependent continuous and discrete design parameters. Recent optical systems, e.g. light-emitting diodes (LEDs, [11, 30, 72]) or vertical-external-cavity surface-emitting-lasers [8, 20], feature multi-layer thin films, which transmit or reflect designated parts of the wave spectrum to achieve a certain functionality [34, 39]. For example, multi-layer thin films are widely used for anti-reflection coating [1, 14, 49, 59, 60]. In general, optimizing those layer stacks with respect to their optical characteristics states an inverse design problem, which covers discrete as well as continuous parameters. Namely, the total number of layers and each layer's dielectric material as well as each layer's thickness. However, considering all these parameters results in a large number of possible designs and particularly in a large number of designs with sub-optimal optical properties. Thus, the corresponding search space is non-convex and contains many sub-optimal local optima [24, 34] of flat fitness [1, 4, 49], which renders gradient-based optimization difficult. Although the existence of a global optimum is mathematically and computationally evinced [14, 59, 60], algorithms that guarantee finding the global optimum in an exhaustive search tend to be computationally intractable, even if only layer thicknesses of less than four layers in total are considered [16]. Thus, in accordance with some theoretical and analytical investigations [1, 15, 28, 68, 71], including genetic and evolutionary approaches, multi-layer thin films are also optimized based on heuristic approaches [10, 21, 29, 41, 44, 67]. Alike many of the mentioned methods, deep learning-assisted techniques are reported to optimize layer thicknesses only: Roberts *et al* [47] proposed a variational autoencoder, Liu *et al* [36] combined forward modeling and inverse design in a tandem of DNNs, and Hegde [23] blended deep learning with evolutionary elements. While the needle-point method (NPM), proposed by Dobrowolski *et al* [55], uses a gradient-based approach to optimize both, materials and thicknesses, of multi-layer thin films, their algorithm cannot fully incorporate dispersive materials, a prerequisite for many optical optimization problems. Note that for comparison we used an implementation of NPM, called OpenFilters [32], to validate some of our results. However, due to the aforementioned non-convexity, solutions found by NPM tend to be insufficient if the initial design differs widely from the optimal one.

In this work, we propose a RL algorithm [62, 63] for the optimization of multi-layer thin films, which is based on multi-path deep Q-learning (MP-DQN, [6]). Our approach allows us to incorporate all three design parameters as well as to operate directly in the space of so-called parameterized actions, where each discrete action is accompanied by a continuous action-parameter. Remarkably, our approach finds designs

from scratch, that is without any pre-determination of the number of layers, supporting the assumption that MP-DQN is able to overcome local optima to some extent. Furthermore, we impose constraints on the design parameters via a Lagrangian formalism, so as to achieve multi-layer thin film designs that feature less complex structures while preserving designated optical characteristics, namely spectral and angular reflectivity, of multi-layer thin films. We demonstrate our algorithm on three different optimization tasks and show that it outperforms multi-layer thin films developed by an expert-guided NPM approach as well as by a standard Q-learning algorithm [29]. In addition, many hyperparameters of MP-DQN are defined such that they have a physical correspondence regarding the proposed multi-layer thin films. Based on this, Q-value estimates are intuitively used to pursue a what-if analysis and thus investigate the behavior of a design under particular layer changes. As such, our approach represents an example of physics-guided explainable artificial intelligence. To achieve discriminability of different multi-layer designs in terms of reward, we introduce a domain-agnostic exponential transformation that can be adapted to other optimization tasks, e.g. when a reconstruction error should be minimized.

2. Parameterized action spaces and inverse design problems: a short research history and merger

RL [57] and especially deep Q-learning have driven major advances in finding an optimal policy in many domains that allow either continuous [35] or discrete actions [43]. The combination of both, discrete and continuous actions, results in parameterized action spaces [42]. Recent work has found sophisticated behavior policies in domains such as 2D robot soccer [6, 22, 26], simulated human–robot interaction [31] and terrain-adaptive bipedal and quadrupedal locomotion [45]. In general, the approaches to solving tasks that include parameterized actions are two-fold: first, hierarchical techniques separate the optimization of discrete actions and continuous action-parameters by iteratively alternating between them during optimization [31, 42]. Therefore, they omit an exchange of information between the policies for discrete and continuous actions, respectively. Second, some recent work focuses on transforming the parameterized actions into continuous [22] or discrete ones [29]. Here, the interaction between continuous and discrete actions is not exploited. Hence, by construction, these concepts are not suitable to represent the intrinsic information contained in parameterized action spaces. However, Xiong *et al* [66] adapted DQN [62] to parameterize each discrete action with a continuous value, thereby incorporating interactions between them. The proposed path-DQN (P-DQN) allows policy optimization directly in a parameterized action space. Bester *et al* [6] suggested so-called MP-DQN, based on their assumption that P-DQN implements the Bellman equation for parameterized action spaces incongruously. Based on MP-DQN, we propose an algorithm for solving inverse design problems that include parameterized actions. Namely, we optimize multi-layer thin films while avoiding unphysical assumptions and sticking closely to the physical domain. For instance, each discrete material choice of a particular layer is parameterized by a continuous thickness value. A sequence of such design choices results in a trajectory of multi-layer thin films with an ascending number of layers in the search space. Conceptionally, this approach can be extended to various other inverse design problems such as the design of meta-lenses [48], which feature continuous (like thicknesses, diameter or angles) and discrete (like materials or basic geometric shapes of components) design parameters.

3. Multi-layer thin films and the inverse design problem

In this work, the design of a multi-layer thin film is specified by three parameters, starting with the total number $L \in \mathbb{N}$ of layers in the layer stack. Each of these consecutive layers consists of a material with a certain refractive index and a specified thickness. Thus, we can encode all parameters of a multi-layer thin film as a vector $\mathbf{n} \in \mathbb{C}^L$ of refractive indices and a vector $\mathbf{t} \in \mathbb{R}^L$ of thickness values, respectively. Based on the transfer matrix method [7] implemented by an open-source Python program [9], the corresponding reflectivity $R_{\lambda,\varphi}(\mathbf{n},\mathbf{t})$ is obtained as a function of the design parameters $\mathbf{n}$ and $\mathbf{t}$ as well as the wavelength λ and the incidence angle φ of the incoming light. Here, a LED functions as a light source that emits an unpolarized electromagnetic spectrum at different angles. We thus get a vector of reflectivity values $\mathbf{R}(\mathbf{n},\mathbf{t}) = (R_{\lambda,\varphi}(\mathbf{n},\mathbf{t}) \,|\, \lambda \in \Lambda, \varphi \in \Phi)$, where $\Lambda, \Phi \subset \mathbb{R}$ denote discrete and compact sets of wavelengths and incidence angles of the emitted radiation, respectively. Based on the intended application of an optical system, the design is required to feature a target reflectivity vector $\mathbf{T} = (T_{\lambda,\varphi} | \lambda \in \Lambda, \varphi \in \Phi)$. Therefore, we can propose an objective function

$$F(\mathbf{n},\mathbf{t},\mathbf{T}) = -\frac{1}{|\Phi| \cdot |\Lambda|} \cdot \sum\nolimits_{\varphi\in\Phi} \sum\nolimits_{\lambda\in\Lambda} |R_{\lambda,\varphi}(\mathbf{n},\mathbf{t}) - T_{\lambda,\varphi}|^2 - \frac{\mu}{L} \cdot \sum\nolimits_{l=1}^{L} t_l \,,\ \mu > 0 \tag{1}$$

that we aim to maximize. Here, the first summand computes the mean squared error between a given and a target reflectivity curve. The multiplier μ in the second addend, a Lagrangian term, introduces regularization, which punishes complex design suggestions. Complexity here refers to the number of layers and layer thicknesses. In principle, the Lagrangian formalism can be used to include additional constraints. However, using this constrained objective function as a reward signal for the RL algorithm results in barely differentiable rewards for designs with reflectivity values close to the target reflectivity. This effect may be attributed to the quadratic form of equation (1), which yields high, but nearly constant values for near-optimal designs. We address this shortcoming by introducing an exponential transformation $r \equiv \exp(\alpha \cdot F)$, $\alpha > 0$, which scales the observed reward r between 0.01 and 1. Here, α is an empirically determined scaling hyperparameter, as explained in appendix B. As

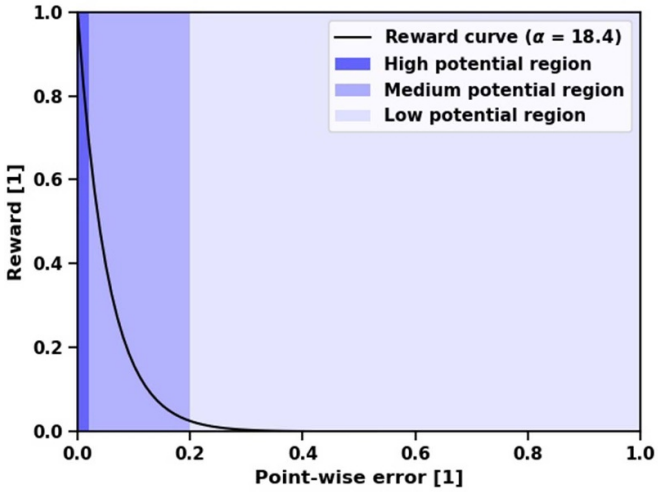


Figure 1. Illustration of the mapping between error and reward, highlighting the regions that divide the search space.

illustrated in figure 1, the reward function now emphasizes the differences in near-optimal system designs while design options with undesirable optical responses are still assigned a low reward. Following the Bellman equation, the discriminability of the rewards is directly imparted to the estimated Q-values, which in turn evaluate the given states. As a result, decision making and learning are improved in general.

4. RL for optimization in parameterized action spaces

In RL, an agent aims to maximize a reward signal that is calculated with respect to the environment's current state. Such a state can be described as a concatenated set $s_i = \{\mathbf{n}, \mathbf{t}\} \subset \mathcal{S}$, where i is the current episode's step number and $\mathcal{S}$ denotes the set of possible states. At the beginning of each of the $E \in \mathbb{N}$ episodes, all entries of the vectors $\mathbf{n}$ and $\mathbf{t}$ are set to zero. As stated in algorithm 1, the agent successively executes parameterized actions $a_i = (n_i, t_i) \in N \times T$, which determine the refractive index $\mathbf{n}_i$ and the thickness $\mathbf{t}_i$ of the current layer $i \leqslant L$. Instead of choosing $\mathbf{n}_i$ and $\mathbf{t}_i$, the agent can also terminate the episode and hereby determine the total number of layers l of the current multi-layer thin film, such that $l \leqslant L$. The parameterized action space becomes $\mathcal{A} = \{a = (n,t) \,|n \in N, t \in T\}$. Obviously, the predefinition of the sets of possible thickness values $T \subset \mathbb{R}^+$ and available refractive indices $N \subset \mathbb{C}$ allows to impose additional hard constraints on the optimization, e.g. to meet manufacturing constraints. Note that our approach is conceptionally applicable in presence of dispersive materials. For convenience and consistency reasons regarding the experiments in section 5, we allowed only real-valued constant refractive indices to be chosen by the agent throughout this study. After an episode is terminated, either by the agent's choice or by reaching the maximum number of layers L, the multi-layer thin film's reflectivity curve is simulated. Based on this reflectivity a reward is assigned, as explained in section 3. In

Algorithm 1 MP-DQN for inverse design optimization

1. Initialize $\theta, \theta', E, L, \mathcal{D}, \tau$
2. **for** $e = 1 : E$ **do**
3. Initialize s_0 (with zeros) and adapt ε
4. **for** $i = 0 : (L-1)$ **do**
5. With probability ε select random action (n_i, t_i)
6. Otherwise select $a_i = (n_i, t_i) = \text{argmax}_{a'}(\widehat{Q}(s_i, a'|\theta))$
7. Stack layer (n_i, t_i) and observe r_i, s_{i+1}
8. Store transition (s_i, a_i, r_i, s_{i+1}) in $\mathcal{D}$
9. **end for**
10. Sample random mini-batch $\mathcal{B} \subset \mathcal{D}$ of transitions $\{(s_j, a_j, r_j, s_{j+1})\}_j$
11. For each transition compute $y = r_j + \gamma \cdot \max_{a'}(\widehat{Q}(s_{j+1}, a'|\theta'))$
12. Compute loss $\mathcal{L} = \sum_{\mathcal{B}}(y - \widehat{Q}(s_j, a_j|\theta))^2$
13. Perform gradient descent on θ following Bester *et al* [6]
14. **if** target network update **then**
15. Update θ' using Polyak averaging $\theta' \leftarrow \tau \cdot \theta + (1-\tau) \cdot \theta'$
16. **end if**
17. **end for**

order to minimize costly calls to the simulation software, each of the non-terminal states is assigned a zero reward. Because these so-called delayed rewards impede Q-value approximation, we rate non-terminal states recursively using an l-step return, $r_{i-1} \leftarrow \gamma \cdot r_i$, $0 < i \leqslant l$, where $r_l \equiv r$ is the final reward and $\gamma = 0.95$ is the discount factor for the future reward.

The described formalism allows us to interpret the problem as a parameterized action Markov decision process $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$ (PAMDP, [42]), where $\mathbb{P}(s_{i+1}|s_i)$ is the Markov state transition probability function. Each transition in this process gets stored in a replay memory $\mathcal{D}$, as a tuple of the current state s_i, the taken action a_i, the subsequent state s_{i+1}, and the l-step return r_i. Using MP-DQN, the collected data and the Bellman equation are used to approximate the Q-values

$$Q(s_i, a_i) = \mathbb{E}_{r_i, s_{i+1}}[r_i + \gamma \cdot \max_{a_{i+1}} Q(s_{i+1}, a_{i+1}) \,|s_i, a_i] \quad (2)$$

that are the expected future rewards given a current state and a particular parameterized action. As a result, the optimal policy $\pi : s \mapsto \text{argmax}_{a'} \widehat{Q}(s, a')$ is given by taking actions a corresponding to maximum Q-value estimates $\widehat{Q}(s,a) \approx Q(s,a)$ in a particular state s. To approximate the Q-values, we implement a sequence of DNNs f and g with joint parameterization θ. Briefly explained, we estimate possible thickness values for each material available given the current state by the network $g : \mathcal{S} \mapsto T^{|N|}$ that features $|s|$ input nodes and $|N|$ output nodes. Each output node corresponds to a material in N and suggests the thickness value of the next layer to stack if the respective material is chosen. Which material is actually chosen is based on the multi-path policy evaluation $f(s, g(s)|\theta)$, with $|N| + 1$ outputs. Each output value represents a Q-value estimate, $\widehat{Q}(s, a|\theta) \equiv \widehat{Q}(s,a)$, for the associated parameterized action while taking into account both, the current state s and the suggestions for thickness values $g(s)$. Note that there is one additional node, which represents the action that terminates an episode. We illustrate our approach

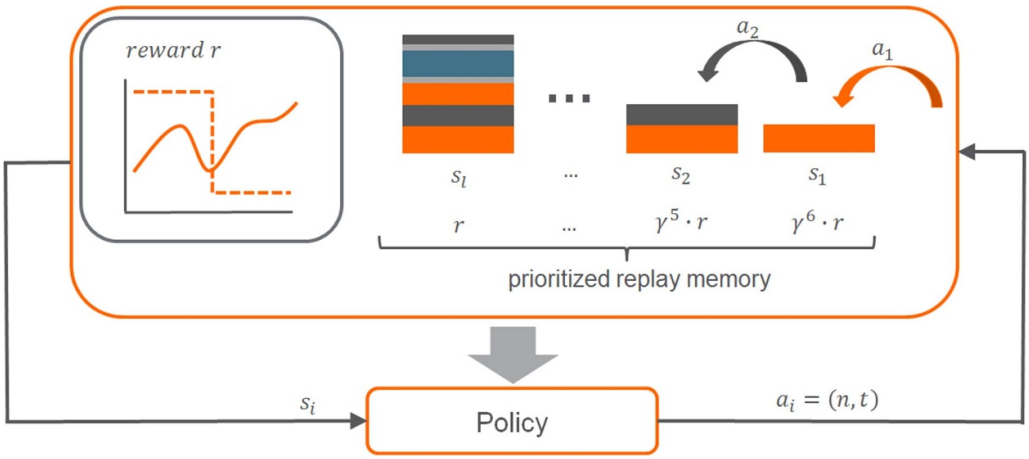


Figure 2. The layer stack is iteratively generated based on the agent's actions. After the terminal state is reached, a simulation reveals the reflectivity behavior over wavelength. This is used to compute the reward signal r and corresponding l-step returns. The obtained experience is stored in a prioritized replay memory and used for adapting the policy.

Table 1. Summary of the tasks including their target curves **T**, considered wavelengths Λ and incident angles Φ. L denotes the maximum number of layers placed, $|N|$ and $|\mathcal{S}|$ are the number of available materials and the approximate number of states of the resulting PAMDP, respectively.

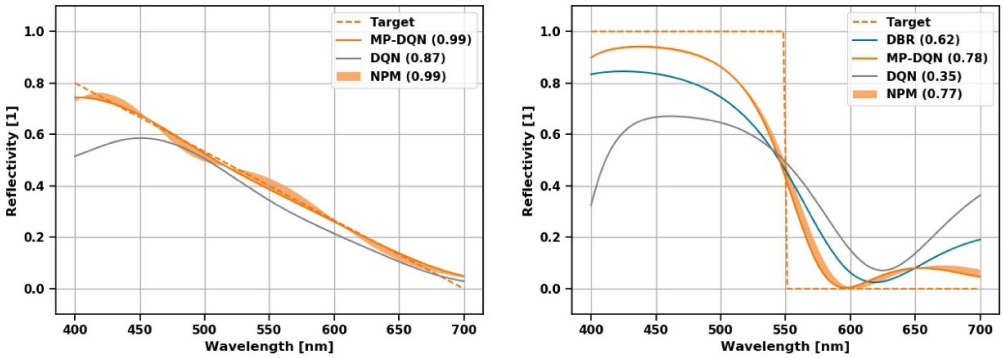
ID	**T**	Λ (nm)	Φ (°)	L	$\lvert\mathcal{S}\rvert$	$\lvert N\rvert$
1	$T_{\lambda,\varphi} = 1/375 \times \lambda - 16/15$	$[400, 700]$	$\{0\}$	8	2.24×10^{29}	4
2	$T_{\lambda,\varphi} = 1/2 \times [1 - \tanh(\lambda - 550)]$	$[400, 700]$	$\{0\}$	8	2.24×10^{29}	4
3	$T_{\lambda,\varphi} = 1.0$	$[445, 455]$	$[0, 60]$	34	1.94×10^{108}	2

in figure 2 which reveals that MP-DQN extends the DQN algorithm so as to solve PAMDPs by considering network g as an intermediate continuous actor and network f as an approximator of Q-values, thus functioning as a discrete actor.

As in common DQNs, the successively collected data is highly correlated and its distribution varies due to policy adaption during optimization. This violates the assumption of independent and identically distributed data for neural network training. Hence, to stabilize policy optimization we introduce a target network [62] and a replay memory $\mathcal{D}$ [43], where sampling from $\mathcal{D}$ breaks the correlation between data generated by the same trajectory. The target and policy networks feature two hidden layers with 256 nodes each. As outlined in algorithm 1, after each episode and entailed l-step return calculation, the policy network parameterization θ is updated with a learning rate of 0.001. The target network parameterization θ' is updated every ten episodes using Polyak averaging, with $\tau = 0.01$. The replay memory was adapted for optical design optimization by implementing a non-uniform random drawing of training batches, so-called prioritization [50]. The probability of choosing a particular transition from the replay memory is determined by applying the softmax function to the losses of transitions. Thus, transitions that correspond to misestimated Q-values have a higher probability of being sampled. Another important aspect of optimization algorithms in general is the exploration-exploitation trade-off that is implemented through an ε-greedy policy in this work. We adapt $\epsilon \in [\epsilon_{\text{final}}, 1]$ before each episode. Beginning from $\varepsilon = 1$, we exponentially reduce ε by a factor of 0.997 until $\epsilon = \epsilon_{\text{final}}$, such that $(1 - \epsilon_{\text{final}})^L \approx 0.3$ holds. This turned out to be an adequate long-term trade-off between exploration and exploitation, as the agent can design a multi-layer thin film in 3 out of 10 episodes without any random exploration. Note that RL is employed to solve an optimization problem. Thus, convergence of the policy is not intended, because this would result in proposing the same optical system again and again without any additional information gain.

5. Experiments

To analyze our MP-DQN approach, we perform optimization on three different tasks, as stated in table 1. For each task, the multi-layer thin film is cladded with air (top side) and a semi-infinite substrate of refractive index $3.194 + 0.00018 \cdot i \in \mathbb{C}$ (bottom side). The choosable materials exhibit real-valued refractive indices. To realize the extent of the corresponding search spaces, we can approximate the total number of possible states to be $|\mathcal{S}| = \sum_{l=1}^{L} |T|^l \cdot |N| \cdot (|N| - 1)^{l-1}$, if we assume discrete layer thicknesses from 0 to 150 nm in steps of 0.1 nm resulting in a total number of $|T| = 1500$ thickness values [29]. We compare our experimental results to multi-layer thin films designed by human experts and another Q-learning algorithm [29], henceforth referred to as DQN algorithm. However, to enhance comparability between our approach and the DQN algorithm, we enabled the latter to not only optimize over layer thicknesses but also over layer materials. Nevertheless, contrary to our approach, DQN operates on discretized thickness values and a pre-defined stack consisting of a fixed number of layers. Therefore, DQN's design initialization was set to a random but fixed layer stack at the beginning of each of the 200 episodes, which cover 250 steps each.



(a) Task 1 (b) Task 2

Figure 3. Illustration of the target and reflectivity curves that correspond to the highest obtained reward using MP-DQN (ours) and DQN. Moreover, we validated our results using NPM. The achieved reward is denoted in brackets. In addition, the reflectivity curve obtained by a distributed Bragg reflector (DBR, see appendix A) is visualized for task 2. We set $\alpha = 18.42$ and $\mu = 0$ in order to compute the reward based on equation (1).

We run DQN ten times and report the reflectivity curves corresponding to the highest achieved rewards for tasks 1 and 2. After running our approach once for 10 000 episodes with L steps each, we compare the results of our approach and the DQN algorithm. Figure 3 reveals that we distinctively outperform DQN not only in terms of achieved best rewards, which were improved by at least 20% for task 1 and 2: Whereas our approach employs 10 000 simulation calls, DQN relies on one simulation call per step resulting in 50 000 simulation calls in each run. Moreover, the same figure states that MP-DQN achieves an even higher reward compared to a distributed Bragg reflector (DBR, see appendix A), which is a physically deduced solution for task 2. The material configuration of the DBR and the multi-layer thin film suggested by our approach are used with randomized layer thickness values as initial designs for NPM to solve task 2. The most suitable reflectivity behavior that was observed in ten NPM runs with randomized initial thicknesses is reported in figure 3. Here, the orange area denotes the deviation between the reflectivities achieved by MP-DQN and NPM. This procedure was repeated for task 1 as well, where we used the design suggested by our approach as initial design for the NPM only. As for both tasks the maximum number $L = 8$ of layers stacked by MP-DQN is fixed, we considered only solutions of NPM consisting of less than ten layers. Notably, we found that NPM did not converge to a sufficient solution if started from scratch for both, task 1 and 2.

5.1. Constrained optimization

To control the complexity of the designs created by our MP-DQN approach, we run task 1 again, using a constrained optimization by setting $\mu = 0.1$ in equation (1). When comparing designs that achieve the same unconstrained reward of approximately 0.99, performing constrained optimization yields a distinctively thinner design with a total thickness of 503.7 nm, whereas the unconstrained approach ($\mu = 0.0$) suggests 598.2 nm. Note that as the constrained reward features an additional non-zero term, the comparison of unconstrained

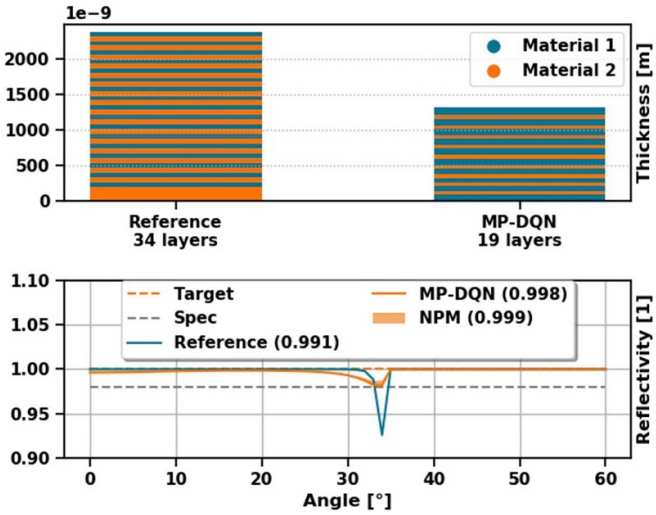


Figure 4. Task 3. On top, the reference design and the design obtained by MP-DQN are depicted. The designs feature alternating layers of two materials with real-valued refractive indices. The bottom illustration depicts the target and specification reflectivity as well as the averaged reflectivities for considered wavelengths over angle. The filled area (orange) represents the solution found by NPM [32] if the design of MP-DQN is used as a starting point.

and constrained reward is invalid. Thus, we report and compare unconstrained rewards for both cases. Due to the convincing proof of concept regarding task 1 and 2, we apply the same Lagrangian multiplier $\mu = 0.1$ to optimize task 3. Due to manufacturing guidelines, only two materials are available for this task. We compare the constrained optimization result with a reference design that consists of 34 layers and was empirically developed by human experts. Namely, optical engineers estimated the number of layers and materials. Afterwards, the aforementioned NPM was used to refine the suggested expert designs. Note that the NPM approach did not converge to a suitable solution if started from scratch. As shown in figure 4, we outperform the reference (Expert + NPM, blue line) and satisfy the specification (Spec, dashed gray line),

Table 2. Each row represents an available material (Mat.), where $\mathrm{Re}(\mathbf{n}_i)$ denotes the real parts of the associated refractive indices. Each column 1–8 corresponds to a layer i. The first sub-row in each column contains the estimated Q-values $\widehat{Q}$ while following the optimal policy for task 2 (bold values). The grayscale values indicate relative differences in the magnitude of Q-values in each column. The second sub-row in each column contains the optical path length p_i, the third sub-row the l-step return r_i resulting if a particular action was taken and we follow the optimal policy in each (other) state.

Mat.	$\mathrm{Re}(\mathbf{n}_i)$	Layer i	1	2	3	4	5	6	7	8
		$\widehat{Q}$	**0.501**	0.297	**0.551**	0.423	**0.631**	0.514	**0.647**	0.509
One	1.457	p_i	0.580	0.380	0.735	0.312	0.790	0.613	1.199	0.376
		r_i	0.544	0.256	0.603	0.429	0.668	0.493	0.741	0.499
		$\widehat{Q}$	**0.388**	0.270	**0.506**	0.484	**0.596**	0.527	0.517	**0.536**
Two	1.645	p_i	0.636	0.834	0.742	0.575	0.939	0.551	0.279	0.357
		r_i	0.414	0.257	0.485	0.477	0.619	0.605	0.568	0.513
		$\widehat{Q}$	0.316	**0.362**	0.416	**0.544**	0.586	**0.578**	**0.612**	**0.714**
Three	1.860	p_i	0.663	0.703	0.967	0.661	1.273	0.609	1.473	0.313
		r_i	0.303	0.337	0.427	0.567	0.566	0.559	0.589	0.780
		$\widehat{Q}$	0.232	**0.539**	0.339	**0.651**	0.457	**0.682**	0.559	0.395
Four	2.327	p_i	0.793	0.694	1.669	0.792	1.647	0.665	1.578	2.296
		r_i	0.182	0.573	0.294	0.634	0.493	0.703	0.575	0.433

using only 19 layers, with 1307.1 nm thickness in total. Practically, this reduction in complexity not only decreases production costs and difficulties but also reduces optical absorption losses in the stack. Finally, to verify our MP-DQN's design (orange line), we set it as a starting point for NPM with randomized thickness values again. The optimization resulted in an increase of the unconstrained reward by 0.001 after ten restarts of NPM with randomized layer thicknesses. The thereby best obtained reflectivity is achieved with 20 layers and illustrated in figure 4 (orange area). Thus, the NPM's marginal improvement is achieved by an undesirable increase of the total thickness by 87.3 nm and—more importantly—one additional layer. This observation indicates that the heuristic MP-DQN finds—at least—a local optimum of the constrained reward regarding layer thicknesses and materials.

5.2. Review from a physical point of view

A physicist's intuition about solving task 2 corresponds to a DBR. Here, our approach coincides with the respective material configuration—except for the last layer. As table 2 shows, the agent places material 3 instead of further alternating between materials 1 and 4. Inspired by the finding that material 4 surprisingly features the lowest Q-value, we analyzed Q-values in terms of optical characteristics. Therefore, we compare the Q-value estimation $\widehat{Q}(s_i, a_i)$ of each transition i of an episode with respect to the optical characteristics of the underlying parameterized action $a_i = (n_i, t_i)$ given the same state s_i. The first optical characteristic that we consider is the refractive index n_i, the second characteristic is the resulting optical path length $p_i = n_i \cdot t_i$. Interestingly, table 2 indicates that the functional dependencies $\widehat{Q}(s_i, n_i) \approx \widehat{Q}(s_i, a_i)$ show monotonic and in general convex behavior and non-convex behavior in case of $\widehat{Q}(s_i, p_i) \approx \widehat{Q}(s_i, a_i)$ for a fixed state s_i, respectively. These relations suggest that the relative order of the Q-value estimates is mainly based on the refractive indices rather than thicknesses that are associated with an action. Moreover, as convexity prohibits the existence of local optima aside from the global optimum, Q-values seem to validly reflect relative adequacy of actions in terms of their associated refractive indices in a particular state. We illustrate this phenomenon in figure 5.

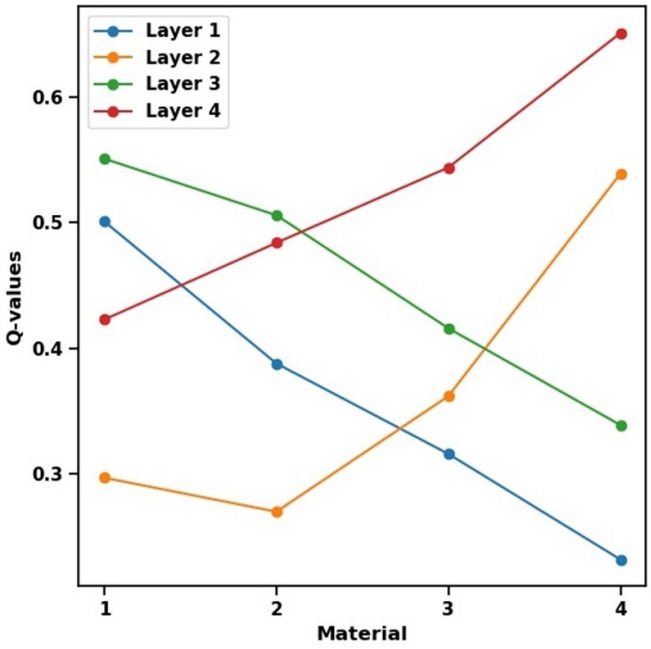


Figure 5. Given task 2, the estimated Q-values corresponding to the first four layers are illustrated. Each Q-value corresponds to an available material. The materials 1–4 are sorted by refractive indices in ascending order.

In addition, the expected future reward enables further physical understanding by conducting a what-if analysis. Namely, Q-values are interpreted as estimations of l-step returns and thus design behavior, e.g. when a particular layer is changed. This was validated by following the optimal policy until layer i, taking a non-optimal parameterized action, and then following the optimal policy again until the terminal state. After conducting this for every possible parameterized

action, the observed l-step returns r_i were collected in table 2. These results indicate that the influence of a design choice on the obtained l-step return is identified by the Q-values. Thus, engineers can infer physical knowledge, e.g. investigating where and why the optimal multi-layer thin film deviates from a physical intuition as exemplified above for task 2. We elucidate the acquired insights about convexity and the what-if analysis in appendix C while also providing information about the learning dynamics.

6. Conclusion

Many inverse design problems in optics and physics feature discrete and continuous parameters, which often makes them only insufficiently solvable. In this work, we present a MP-DQN framework that presents a heuristic solution to inverse design problems. We demonstrated the suitability of our approach using the example of multi-layer thin film optimization, which includes discrete as well as continuous parameters. The environment for the RL agent is emulated by a simulation, based on whose outcomes the reward is computed. Thus, the proposed method can be used in absence of gradient information or prior assumptions about the system under consideration as well. Our contribution is three-fold: first, we demonstrate how to formulate inverse design problems as parameterized Markov decision processes in order to solve them with parameterized RL. Notably, our approach abandons the unphysical reduction of the search space as well as the need to rely on prior beliefs about the underlying system, which both may lead to sub-optimal results. Hence, system designs are optimized with respect to their entire physical structure and as a result, our approach distinctively outperforms other methods. Second, we develop a general constrained objective function to compute rewards based on an exponential transformation. The resulting reward signal becomes differentiable, which eases the agent's policy optimization and decision making. Moreover, it enables us to control the complexity of suggested system designs, which reduces production costs and decreases optical absorption losses in multi-layer thin films. Finally, we perform a what-if analysis based on Q-value estimates and thereby demonstrate how physicists can gain insights from the estimated Q-values. Eventually, the proposed approach represents an example of physics-guided explainable artificial intelligence.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

This research was conducted in collaboration with OSRAM Opto Semiconductors GmbH. We thank our dear colleagues for their assistance regarding optical expertise and editorial input, especially Daniel Grünbaum for various comments that greatly improved this manuscript.

The authors declare no competing interests.

Appendix A. Filter construction using distributed Bragg reflectors

A distributed Bragg reflector (DBR) is an efficient optical reflector that consists of alternating thin films of materials with different refractive indices. Basically, a DBR is determined by two thickness values $t_1, t_2 \in \mathbb{R}^+$ and real refractive indices $n_1, n_2 \in \mathbb{R}^+$, where $n_1 < n_2$ holds. Task 2 of table 1 corresponds to a high-pass filter in the wavelength domain, because wavelengths lower than 550 nm should be reflected. To obtain a physically deduced filter and thus a solution to task 2, we can use a DBR [39]. Here, the wavelength width $\Delta\lambda$ of the stopping band can be computed with respect to the center wavelength λ_0 of the stopping band. In addition, we want the stopping band to end at 550 nm and set $n_1 = 1.457$ and $n_2 = 2.327$. The obtained linear equation system

$$\begin{aligned} \Delta\lambda &= \frac{4}{\pi} \cdot \lambda_0 \cdot \arcsin\left|\frac{n_2 - n_1}{n_2 + n_1}\right| \\ \lambda_0 + \Delta\lambda &= 550\,\text{nm} \end{aligned}$$

yields $\lambda_0 = 424.59$ nm. The resonance condition for first order constructive interference $n_1 \cdot t_1 = n_2 \cdot t_2 = \lambda_0/4$ yields $t_1 = 72.85$ nm and $t_2 = 45.62$ nm. To obtain an eight-layer DBR of 473.88 nm total thickness, we repeatedly stack these two layers four times.

Appendix B. Impact of reward transformation and Q-value reliability

Following the optimal policy, which leads to an optimal multi-layer thin film, relies on an accurate Q-value estimation for as many state-action pairs as possible. Moreover, to ease decision making, the Q-value estimates for particular parameterized actions should be as distinguishable as possible. This condition does not apply if the rewards of more and more improved designs remain almost constant, because equation (2) and its implementation in algorithm 1 reveal that in such a case the Q-value estimates will be almost constant, too. On the other hand, many regions in the design search space are completely inadequate for solving a given task and should be assigned a very small reward. This is why we introduce a dedicated reward transformation, which relies on a hyperparameter $\alpha > 0$ and is illustrated in figure 1. The hyperparameter is computed by

$$\alpha = -\frac{1}{\eta} \cdot \ln\left(\frac{\beta_1}{\beta_2}\right) = 18.42,$$

where $\beta_1 = 0.01$ and $\beta_2 = 1.0$ are the lower and upper bound hyperparameters for the reward, respectively. The empirical mean value $\eta = 0.25$ of equation (1) is computed based on

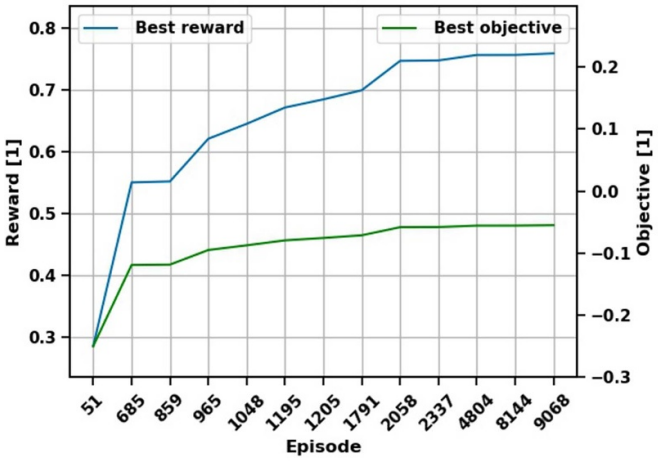


Figure 6. The best obtained objective (1) and reward over episode of its achievement for $\alpha = 18.42$ reveals the higher discriminability of designs during training. Axis limits are chosen such that the absolute length of the axes of reward and error coincide.

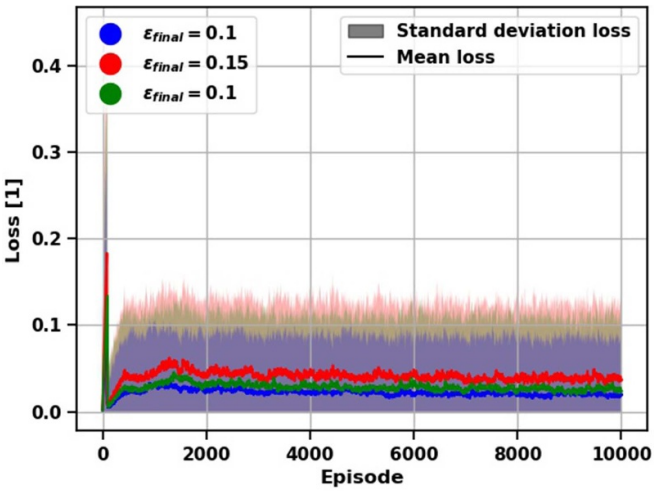


Figure 7. Illustration of the standard deviation and mean value of the computed loss over episode. We investigated different configurations of ϵ_{final}. Note that we omitted the loss-weighted sampling of mini-batches in one case (green).

1000 randomly drawn multi-layer thin films. The impact of this transformation regarding task 2 is illustrated in figure 6.

It is often not discussed that the approximation of Q-values can be monitored during policy optimization. In figure 7, we depict the mean value and standard deviation of the loss $\mathcal{L}$ for task 2, which is computed every episode according to algorithm 1, based on the entire data in the replay memory $\mathcal{D}$. Unsurprisingly, in the beginning, the loss is high, because the training, which is based on batches of size 128, starts when the replay memory of total size 5000 contains an initial number of 500 transitions. This prevents the neural network parameterizations from being biased due to very limited data in the early training phase. Moreover, the impact of different final exploration probabilities ϵ_{final} and the effect of prioritization is observable. Whereas a higher value for ϵ_{final} implies more exploration of unknown regions of the search space and thus uncertainty in the underlying Q-value estimation, prioritization reduces the standard deviation of loss values by preferring misestimated transitions for sampling into the mini-batches used for training. Monitoring the approximation of Q-values in the replay memory can function as an indicator in many respects. For example to answer the questions of whether to initiate more exploration in case of overfitting or whether the engineers can trust a Q-value approximation in general or should adapt their hyperparameters.

Appendix C. Learning dynamics

In addition to the loss, we also tracked l-step returns and eventually achieved rewards for each episode. Figure 8 depicts these measures for two different values of the final exploration probability ϵ_{final} solving task 2. As expected, we achieve higher running rewards with lower ϵ_{final}. In addition and more importantly, the best obtained reward remains nearly stable in both cases although the best proposed optical design differs due to various local optima in the search space. Finally, we investigated how the functional behavior of Q-values evolves

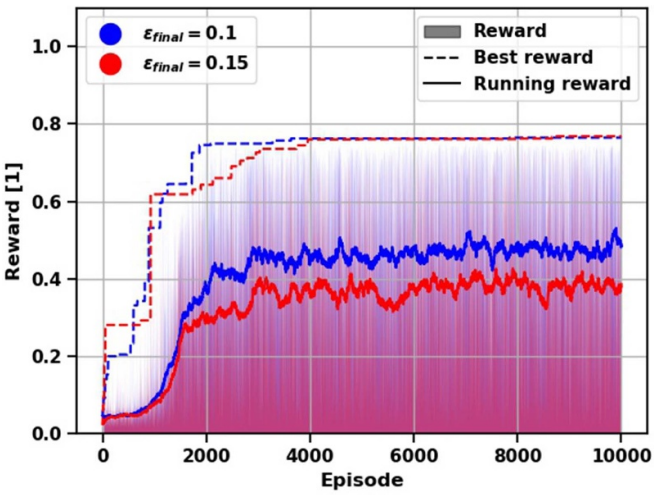


Figure 8. Illustration of obtained reward (filled area) and running reward (solid line) over episode. The best obtained reward is indicated by dashed lines for two configurations of ϵ_{final}.

during optimization. As a Q-value is related to a parameterized action in step i, we characterize the latter by either the refractive index n_i or the optical path length p_i. We track in each step $i \leqslant L$ of an episode whether the estimated Q-values are convex in terms of refractive index $\widehat{Q}(s_i, n_i) \approx \widehat{Q}(s_i, a_i)$ or optical path length $\widehat{Q}(s_i, p_i) \approx \widehat{Q}(s_i, a_i)$ given the same state s_i. Based on the tracked data, the ratio between convex estimates and the total number of steps in each episode is calculated. Figure 9 illustrates how the running mean and standard deviation of these ratios evolve over episodes. Here, an additional measure is covered: The ratio of steps in each episode that were convex with respect to both, refractive index and optical path length. As we estimate four material-related Q-values per step, the combinatorially deduced probability for the estimates to show convex behavior is 50%. This regime of randomness is indicated by the black rule in figure 9. The running mean and

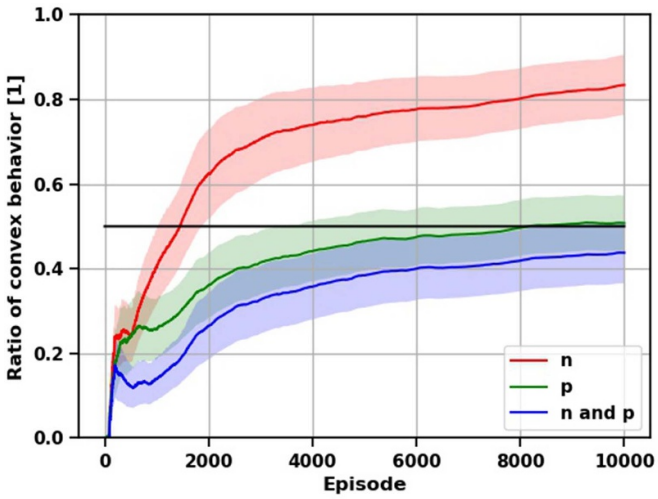


Figure 9. Step ratio of convex behavior of Q-value approximation in terms of refractive index (n), optical path length (p), and corresponding coincidence (n and p) over episode, respectively.

standard deviation of ratios were computed based on Welford's online algorithm [64]. Although the optical path length intuitively gives a more encompassing optical information about a parameterized action, the ratios of convex behavior based on refractive indices (red rule) of 0.6–0.8 are higher than for optical path lengths that are around random guessing at 0.5–0.6. Moreover, a comparison indicates that if the Q-value estimates are convex in terms of optical path length (green rule), they are also convex in terms of refractive indices and thus both optical characteristics (blue rule). In general, coincident convexity in terms of both optical characteristics cannot be proven. But it seems that the Q-value estimates reflect some optical characteristics and thus contain information about the optical similarity of corresponding parameterized actions.

ORCID iDs

Heribert Wankerl https://orcid.org/0000-0001-5634-4038
Maike L Stern https://orcid.org/0000-0001-9989-5868
Elmar W Lang https://orcid.org/0000-0001-7440-0224

References

[1] Anzengruber S W, Klann E, Ramlau R and Tonova D 2012 Numerical methods for the design of gradient-index optical coatings *Appl. Opt.* **51** 8277–95
[2] Banerji S, Majumder A, Hamrick A, Menon R and Sensale-Rodriguez B 2020 Machine learning enables design of on-chip integrated silicon t-junctions with footprint of 1.2 micrometer × 1.2 micrometer *Nano Commun. Netw.* **25** 100312
[3] Banerji S, Majumder A, Hamrick A, Menon R and Sensale-Rodriguez B 2021 Ultra-compact integrated photonic devices enabled by machine learning and digital metamaterials *OSA Continuum* **4** 602–7
[4] Becker H, Tonova D, Sundermann M, Ehlers H, Günster S and Ristau D 2014 Design and realization of advanced multi-index systems *Appl. Opt.* **53** A88–95
[5] Bellman R 2003 *Dynamic Programming* (New York: Dover Publications Inc)
[6] Bester C J, James S D and Konidaris G D 2019 Multi-pass q-networks for deep reinforcement learning with parameterised action spaces (arXiv:1905.04388v1)
[7] Born M and Wolf E 1959 *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Oxford: Pergamon)
[8] Broda A, Kuźmicz A, Rychlik G, Chmielewski K, Wójcik-Jedlińska A, Sankowska I, Gołaszewska-Malec K, Michalak K and Muszalski J 2017 Highly efficient heat extraction by double diamond heat-spreaders applied to a vertical external cavity surface-emitting laser *Opt. Quantum Electron.* **49** 287
[9] Byrnes S J 2018 Multilayer optical calculations (arXiv:1603.02720v5)
[10] Chang C P, Lee Y H and Wu S Y 1990 Optimization of a thin-film multilayer design by use of the generalized simulated-annealing method *Opt. Lett.* **15** 595–7
[11] Chen C, Chang S, Su Y, Chi G, Sheu J and Chen J 2002 High-efficiency InGaN-GaN MQW green light-emitting diodes with CART and DBR structures *IEEE J. Sel. Top. Quantum Electron.* **8** 284–8
[12] Cybenko G 1989 Approximation by superpositions of a sigmoidal function *Math. Control Sig. Syst.* **2** 303–14
[13] Cybenko G 1989 Multilayer feedforward networks are universal approximators *Neural Netw.* **2** 359–66
[14] Dobrowolski J A, Tikhonravov A V, Trubetskov M K, Sullivan B T and Verly P G 1996 Optimal single-band normal-incidence antireflection coatings *Appl. Opt.* **35** 644–58
[15] Ebrahimi M and Ghasemi M 2018 Design and optimization of thin film polarizer at the wavelength of 1540 nm using differential evolution algorithm *Opt. Quantum Electron.* **50** 1–9
[16] Azunre P *et al* 2019 Guaranteed global optimization of thin-film optical systems *New J. Phys.* **21** 073050
[17] Gal Y and Ghahramani Z 2016 Dropout as a Bayesian approximation: representing model uncertainty in deep learning *Proc. 33rd Int. Conf. on Machine Learning (PMLR)* vol **48**, 1050–9
[18] Gandhi A and Png C E 2019 Modal classification in optical waveguides using deep learning *J. Mod. Opt.* **66** 557–61
[19] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)
[20] Guina1 M, Rantamäki1 A and Härkönen1 A 2017 Optically pumped VECSELs: review of technology and progress *J. Phys. D: Appl. Phys.* **50** 383001
[21] Guo X, Zhou H Y, Guo S, Luan X X, Cui W K, Ma Y F and Shi L 2014 Design of broadband omnidirectional antireflection coatings using ant colony algorithm *Opt. Express* **22** A1137–44
[22] Hausknecht M and Stone P 2016 Deep reinforcement learning in parameterized action space *Proc. Int. Conf. on Learning Representations* (*San Juan, Puerto Rico*)
[23] Hedge R S 2019 Accelerating optics design optimizations with deep learning *Opt. Eng., Bellingham* **58** 065103
[24] Horst R and Tuy H 1996 *Global Optimization: Deterministic Approaches* (Berlin: Springer)
[25] Huang Z, Liuab X and Zang J 2019 The inverse design of structural color using machine learning *Nanoscale* **11** 21748–58
[26] Hussein A, Elyan E and Jayne C 2018 Deep imitation learning with memory for robocup soccer simulation *Proc. Int. Conf. Eng. Appl. Neural Netw.* (*Bristol, UK, 3-5 September 2018*) 31–43
[27] Inampudi S and Mosallaei H 2018 Neural network based design of metagratings *Appl. Phys. Lett.* **112** 241102

[28] Janicki V, Sancho-Parramon J and Zorc H 2008 Refractive index profile modelling of dielectric inhomogeneous coatings using effective medium theories *Thin Solid Films* **516** 3368–73

[29] Jiang A, Osamu Y and Chen L 2020 Multilayer optical thin film design with deep q learning *Sci. Rep.* **10** 12780

[30] Khadir S, Chakaroun M, Belkhir A, Fischer A, Lamrous O and Boudrioua A 2015 Localized surface plasmon enhanced emission of organic light emitting diode coupled to DBR-cathode microcavity by using silver nanoclusters *Opt. Express* **23** 23647–59

[31] Khamassi M, Velentzas G, Tsitsimis T and Tzafestas C 2017 Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task *Proc. First IEEE Int. Conf. Robot. Comput.* 28–35

[32] Larouche S and Martinu L 2008 OpenFilters: open-source software for the design, optimization and synthesis of optical filters *Appl. Opt.* **47** C219–30

[33] Lecun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–444

[34] Liddell H M and Jerrard H G 1982 Computer-aided techniques for the design of multilayer filters *Opt. Laser Tech.* **14** 51

[35] Lillicrap T P, Hunt J J, Pritzel A, Heess N, Ereza T, Tassa Y, Silver D and Wierstra D 2015 Continuous control with deep reinforcement learning (arXiv:1509.02971)

[36] Liu D, Tan Y, Khoram E and Yu Z 2018 Training deep neural networks for the inverse design of nanophotonic structures *ACS Photonics* **5** 1365–9

[37] Liu Z, Zhu D, Rodrigues S P, Lee K-T and Cai W 2018 Generative model for the inverse design of metasurfaces *Nano Lett.* **18** 6570–6

[38] Ma W, Cheng F and Liu Y 2018 Deep-learning-enabled on-demand design of chiral metamaterials *ACS Nano* **12** 6326–34

[39] MacLeod H A and Macleod H A 2010 *Thin-Film Optical Filters* (Boca Raton, FL: CRC Press)

[40] Malkiel I, Mrejen M, Nagler A, Arieli U, Wolf L and Suchowski H 2018 Plasmonic nanostructure design and characterization via deep learning *Light: Sci. Appl.* **7** 60

[41] Martin S, Rivory J and Schoenauer M 1995 Synthesis of optical multilayer systems using genetic algorithms *Appl. Opt.* **34** 2247–54

[42] Masson W, Ranchod P and Konidaris G 2016 Reinforcement learning with parameterized actions *Proc. Thirtieth Conf. on Artificial Intelligence* pp 1934–40

[43] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D and Riedmiller M 2013 Playing Atari with deep reinforcement learning (arXiv:1312.5602v1)

[44] Paszkowicz W 2013 Genetic algorithms, a nature-inspired tool: a survey of applications in materials science and related fields: part II *Mater. Manuf. Process.* **28** 708–25

[45] Peng X B, Berseth G and de Panne M V 2016 Terrain-adaptive locomotion skills using deep reinforcement learning *ACM Trans. Graph.* **35** 81

[46] Peurifoy J, Shen Y, Jing L, Yang Y, Cano-Renteria F, DeLacy B G, Joannopoulos J D, Tegmark M and Soljačić M 2018 Nanophotonic particle simulation and inverse design using artificial neural networks *Sci. Adv.* **4** 6

[47] Roberts J and Wang E W 2018 Modeling and optimization of thin-film optical devices using a variational autoencoder *212 CS229 Final Project Report Fall 2018* (Stanford University)

[48] Sajedian I, Badloe T and Rho J 2019 Optimisation of colour generation from dielectric nanostructures using reinforcement learning *Opt. Express* **27** 5874–83

[49] Schallenberg U B 2006 Antireflection design concepts with equivalent layers *Appl. Opt.* **45** 1507–14

[50] Schaul T, Quan J, Antonoglou I and Silver D 2016 Prioritized experience replay *Proc. Int. Conf. on Learning Representations*

[51] Schrittwieser J *et al* 2020 Mastering Atari, Go, chess and shogi by planning with a learned model *Nature* **588** 604–9

[52] Schwartz-Ziv R and Tishby N 2017 Opening the black box of deep neural networks via information (arXiv:1703.00810v3)

[53] Silver D *et al* 2016 Mastering the game of go with deep neural networks and tree search *Nature* **529** 484–9

[54] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guezv A, Lanctot M, Sifre L, Kumaran D and Graepel T 2017 Mastering chess and shogi by self-play with a general reinforcement learning algorithm (arXiv:1712.01815v1)

[55] Sullivan B T and Dobrowolski J A 1996 Implementation of a numerical needle method for thin-film design *Appl. Opt.* **35** 5484–92

[56] Sun C, Kaiser E, Brunton S L and Kutz J N 2020 Deep reinforcement learning for optical systems: a case study of mode-locked lasers (arXiv:2006.05579v1)

[57] Sutton R S and Barto A 1998 *Introduction to Reinforcement Learning* (Cambridge: MIT Press)

[58] Tahersima M H, Kojima K, Koike-Akino T, Jha D, Wanga B, Lin C and Parsons K 2019 Deep neural network inverse design of integrated photonic power splitters *Sci. Rep.***9** 1368

[59] Tikhonravov A V 1993 Some theoretical aspects of thin-film optics and their applications *Appl. Opt.* **32** 5417–26

[60] Tikhonravov A V and Dobrowolski J A 1993 Quasi-optimal synthesis for antireflection coatings: a new method *Appl. Opt.* **32** 4265–75

[61] Tünnermann H and Shirakawa A 2019 Deep reinforcement learning for coherent beam combining applications *Opt. Express* **27** 24223–30

[62] van Hasselt H, Guez A and Silver D 2016 Deep reinforcement learning with double q-learning *Proc. Thirteenth Conf. on Artificial Intelligence Phoenix, Arizona* 2094–100

[63] Watkins C 1989 Learning from delayed rewards PhD Thesis

[64] Welford B P 1962 Note on a method for calculating corrected sums of squares and products *Technometrics* **4** 419–20

[65] Wiecha P R, Lecestre A, Mallet N and Larrieu G 2019 Pushing the limits of optical information storage using deep learning *Nat. Nanotechnol.* **14** 237–44

[66] Xiong J, Qing Wang Z Y, Sun P, Lei Han Y Z, Fu H, Zhang T and Ji Liu H L 2018 Parametrized deep q-networks learning: reinforcement learning with discrete-continuous hybrid action space (arXiv:1810.06394v1)

[67] Yang C, Hong L, Shen W, Zhang Y, Liu X and Zhen H 2013 Design of reflective color filters with high angular tolerance by particle swarm optimization method *Opt. Express* **21** 9315–23

[68] Yang J-M and Kao C-Y 2001 Efficient evolutionary algorithm for the thin-film synthesis of inhomogeneous optical coatings *Appl. Opt.* **40** 3256–67

[69] Yao K, Unni R and Zheng Y 2019 Intelligent nanophotonics: merging photonics and artificial intelligence at the nanoscale *Nanophotonics* **8** 339–66

[70] Zhang T, Wang J, Liu Q, Zhou J, Dai J, Han X, Zhou Y and Xu K 2019 Spectrum prediction and inverse design for plasmonic waveguide system based on artificial neural networks *Photon. Res.* **7** 368–80

[71] Zhao Y, Chen F, Shen Q and Zhang L 2014 Design of reflective color filters with high angular tolerance by particle swarm optimization method *Prog. Electromagn. Res.* **145** 39–45

[72] Zhou S, Liu X, Gao Y, Liu Y, Liu M, LIU Z, Gui C and Liu S 2017 Numerical and experimental investigation of GaN-based flip-chip light-emitting diodes with highly reflective Ag/TiW and ITO/DBR ohmic contacts *Opt. Express* **25** 26615–27