Parameterized Reinforcement Learning for Optical System Optimization

Heribert Wankerl
University of Regensburg
OSRAM Opto Semiconductors
Regensburg, 93053
heribert.wankerl@osram-os.com

Maike L. Stern
OSRAM Opto Semiconductors
Regensburg, 93053

Ali Mahdavi
OSRAM Opto Semiconductors
Regensburg, 93053

Christoph Eichler
OSRAM Opto Semiconductors
Regensburg, 93053

Elmar W. Lang
University of Regensburg
Regensburg, 93053

Abstract

Designing a multi-layer optical system with designated optical characteristics is an inverse design problem in which the resulting design is determined by several discrete and continuous parameters. In particular, we consider three design parameters to describe a multi-layer stack: Each layer's dielectric material and thickness as well as the total number of layers. Such a combination of both, discrete and continuous parameters is a challenging optimization problem that often requires a computationally expensive search for an optimal system design. Hence, most methods merely determine the optimal thicknesses of the system's layers. To incorporate layer material and the total number of layers as well, we propose a method that considers the stacking of consecutive layers as parameterized actions in a Markov decision process. We propose an exponentially transformed reward signal that eases policy optimization and adapt a recent variant of Q-learning for inverse design optimization. We demonstrate that our method outperforms human experts and a naive reinforcement learning algorithm concerning the achieved optical characteristics. Moreover, the learned Q-values contain information about the optical properties of multi-layer optical systems, thereby allowing physical interpretation or what-if analysis.

1 Introduction

Modern optical systems feature complex multi-layer designs, which transmit or reflect designated parts of the wave spectrum to achieve a certain functionality [9, 11]. Optimizing those layer stacks with respect to their optical characteristics is an inverse design problem, which covers discrete as well as continuous parameters. Namely, the total number of layers and each layer's dielectric material properties as well as each layer's thickness. However, considering all these parameters results in a large number of possible designs and more 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 [9, 5]. As a result, this kind of optimization problem is often solved with heuristic approaches that only optimize one parameter, as an instance the layers' thicknesses [2, 15, 25, 3, 12], or transform the search space by considering only the discretized layer thickness values [7]. While Dobrowolski et al. [19] allow to incorporate discrete and continuous parameters, their algorithm cannot incorporate dispersive materials, a prerequisite for many optical optimization problems. Other recent approaches require the pre-selection of an extensive dataset to train a differentiable surrogate

Workshop on machine learning for engineering modeling, simulation and design @ NeurIPS 2020

model in a supervised manner [17]. In this work, we propose a reinforcement learning algorithm (RL, [22, 21]) for the optimization of multi-layer optical systems, which is based on multi-path deep Q-learning (MP-DQN, [1]). Our approach allows us to incorporate all three design parameters and to operate directly in the space of so-called parameterized actions, where each discrete action is accompanied by a continuous action-parameter. Furthermore, we impose constraints on the design parameters via a Lagrangian formalism, so as to achieve a system design that features less complex structures while preserving designated reflectivity characteristics. We demonstrate our algorithm on three different optimization tasks and show that it outperforms optical system designs developed by human experts as well as a standard Q-learning algorithm [7]. In addition, many hyperparameters of MP-DQN are defined such that they have a physical correspondence regarding the proposed optical systems. 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.

2 State of the art

RL [20] and especially deep Q-learning have driven major advances in finding an optimal policy in many domains that allow either continuous actions [10] or discrete actions [14]. The combination of both, discrete and continuous actions results in parameterized action spaces [13]. Recent work has found sophisticated behavior policies in domains such as 2D robot soccer [4, 6, 1], simulated human-robot interaction [8] and terrain-adaptive bipedal and quadrupedal locomotion [16]. 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 [13, 8]. 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 [4] or discrete ones [7]. 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. [24] adapted deep Q-learning (DQN, [21]) 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. [1] suggested so-called multi-path DQN (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 optical systems while avoiding unphysical assumptions and sticking closely to the physical domain. For instance, each discrete material choice is parameterized by a continuous thickness value. A sequence of such design choices results in a multi-layer optical system. To achieve discriminability of different optical systems 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.

3 Optical systems and the inverse design problem

In this work, the design of an optical system 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 layer as a vector $\mathbf{n} \in \mathbb{C}^L$ of refractive indexes and a vector $\mathbf{t} \in \mathbb{R}^L$ of thickness values, respectively. Based on a simulation, the observed 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 incident angle φ of the incoming light. Here, a light-emitting diode 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

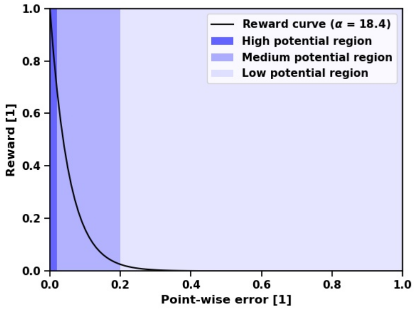


Figure 1: Illustration of the mapping between error and reward, highlighting the regions that divide the search space.

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

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

that we aim to maximize. Here, the first summand computes the mean squared error (MSE) 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. However, using this constrained object 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 $r \equiv \exp\left(\alpha \cdot F\right), \alpha > 0$, 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 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 Reinforcement learning 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 $i \leq L$. 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 optical system, such that $l \leq L$. The parameterized action space becomes $\mathcal{A} = \{a = (n, t) \,|n \in N, t \in T\}$. Obviously, the pre-definition of the sets of possible thickness values $T \subset \mathbb{R}^+$ and available refractive indexes $N \subset \mathbb{C}$ allows to impose additional hard constraints on the optimization. After an episode is terminated, either by the agent's choice or by reaching the maximum number of layers L, the optical system's reflectivity curve is simulated. Based on this reflectivity a reward is assigned, as explained in section 3. In order to minimize costly calls to the simulation software, each of the non-terminal states

Algorithm 1 MP-DQN for inverse design optimization

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

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, $r_{i-1} \leftarrow \gamma \cdot r_i, \, , 0 < i \leq l$, 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, [13]), where $\mathbb{P}\left(s_{i+1}|s_i\right)$ 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

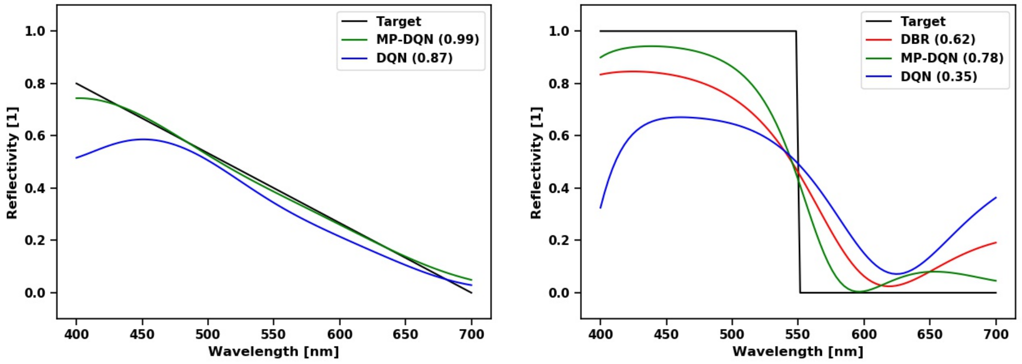
$$Q(s_i, a_i) = \mathbb{E}_{r_i, s_{i+1}}[r_i + \gamma \cdot \max_{a_{i+1}} Q\left(s_{i+1}, a_{i+1}\right)|s_i, a_i] \tag{2}$$

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 deep neural networks 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 can summarize 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 [21] and a replay memory $\mathcal{D}$ [14], where sampling from $\mathcal{D}$ breaks the correlation between data generated by the same trajectory. The target and policy network 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 [18]. 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 getting 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 $\epsilon \in [\epsilon_{final}, 1]$ before each episode. Beginning from $\epsilon = 1$, we exponentially reduce ϵ by a factor of 0.997 until $\epsilon = \epsilon_{final}$, such that $(1 - \epsilon_{final})^L \approx 0.3$ holds. This turned out to be an adequate long-term trade-off between exploration and exploitation as the agent can design an optical system 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. To realize the extend 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 $0 - 150\,nm$ in steps of $0.1\,nm$ resulting in a total number of $|T| = 1500$ thickness values [7]. We compare our experimental results to optical systems designed by human experts and another Q-learning algorithm [7], 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 for layer material. 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



(a) Task 1

(b) Task 2

Figure 2: Illustration of the target and reflectivity curves that correspond to the highest obtained reward using MP-DQN (ours) and DQN. 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).

beginning of each of the 200 episodes, which cover 250 steps each. 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 2 reveals that we distinctively outperform DQN not only in terms of achieved best rewards that 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.

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 and constrained reward is invalid. Thus, we report and compare unconstrained rewards for both cases. Due to the convincing results, we apply the same Lagrangian multiplier $\mu = 0.1$ to optimize task 3. We compare the constrained optimization result with a reference design that consists of 34 layers and was developed by human experts. As shown in figure 3, we outperform the reference and satisfy the specification (Spec, red line), using only 19 layers, with $1307.1\,nm$ thickness in total. Practically, this reduction in complexity not only decreases production costs but also reduces optical absorption losses in the stack.

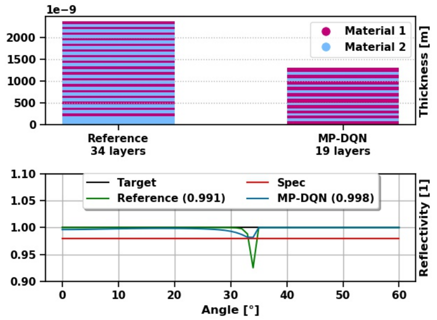


Figure 3: Task 3. On top, the reference design and the design obtained by MP-DQN is depicted. The bottom illustration depicts the target and specification reflectivity as well as the averaged reflectivities for considered wavelengths over angle.

ID	$\mathbf{T}$	$\Lambda[nm]$	$\Phi[°]$	L	$\|\mathcal{S}\|$	$\|N\|$
1	$T_{\lambda,\varphi} = {}^{1}/_{375} \cdot \lambda - {}^{16}/_{15}$	$[400, 700]$	$\{0\}$	8	$2.24 \cdot 10^{29}$	4
2	$T_{\lambda,\varphi} = {}^{1}/_{2} \cdot [1 - \tanh(\lambda - 550)]$	$[400, 700]$	$\{0\}$	8	$2.24 \cdot 10^{29}$	4
3	$T_{\lambda,\varphi} = 1.0$	$[445, 455]$	$[0, 60]$	34	$1.94 \cdot 10^{108}$	2

Table 1: Summary of the tasks including their target curves $\mathbf{T}$, 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.

Mat.	$\mathrm{Re}(\mathbf{n}_i)$	Layer i	1	2	3	4	5	6	7	8
1	1.457	$\widehat{Q}$	**0.501**	**0.297**	**0.551**	**0.423**	**0.631**	**0.514**	**0.647**	**0.509**
		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
2	1.645	$\widehat{Q}$	**0.388**	**0.270**	**0.506**	**0.484**	**0.596**	**0.527**	**0.517**	**0.536**
		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
3	1.860	$\widehat{Q}$	**0.316**	**0.362**	**0.416**	**0.544**	**0.586**	**0.578**	**0.612**	**0.714**
		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
4	2.327	$\widehat{Q}$	**0.232**	**0.539**	**0.339**	**0.651**	**0.457**	**0.682**	**0.559**	**0.395**
		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

Table 2: Each row represents an available material (Mat.), where $\mathrm{Re}(\mathbf{n}_i)$ denotes the real parts of the associated refractive indexes. 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. 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.

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 $\widehat{Q}(s_i, n_i) \approx \widehat{Q}(s_i, a_i)$ shows 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 indexes 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 indexes in a particular state.

In addition, the expected future reward provides 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 optical system 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.

Conclusion

In this work, we introduce a novel method to optimize optical system designs that require discrete as well as continuous parameters, using multi-path deep Q-learning (MP-DQN). Our contributions are three-fold: First, we used MP-DQN to address constrained inverse design problems, by formulating them as parameterized action Markov decision processes. Notably, our approach abandons the unphysical discretization of continuous variables and as a result, distinctively outperforms other methods. Second, we developed a 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 the system designs, which reduces production costs and decreases optical absorption losses. Finally, we performed a what-if analysis based on Q-value estimates and thereby demonstrate how optical engineers can gain physical insights from the estimated Q-values.

Acknowledgments and Disclosure of Funding

This research was conducted at OSRAM Opto Semiconductors GmbH.
We thank our 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.

References

[1] C. J. Bester, S. D. James, and G. D. Konidaris. Multi-pass q-networks for deep reinforcement learning with parameterised action spaces. *arXiv preprint*, 2019.

[2] C. P. Chang, Y. H. Lee, and S. Y. Wu. Optimization of a thin-film multilayer design by use of the generalized simulated-annealing method. *Optics Letters*, 1990.

[3] X. Guo, H. Y. Zhou, S. Guo, X. X. Luan, W. K. Cui, Y. F. Ma, , and L. Shi. Design of broadband omnidirectional antireflection coatings using ant colony algorithm. *Optics Express*, 22, 2014.

[4] M. Hausknecht and P. Stone. Deep reinforcement learning in parameterized action space. *In: Proceedings of the International Conference on Learning Representations*, 2016.

[5] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer, 1996.

[6] A. Hussein, E. Elyan, and C. Jayne. Deep imitation learning with memory for robocup soccer simulation. *In Proceedings of the International Conference on Engineering Applications of Neural Networks, Springer*, 2018.

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

[8] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas. Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task. *In Proceedings of the First IEEE International Conference on Robotic Computing*, 2017.

[9] H. M. Liddell and H. G. Jerrard. *Computer-Aided Techniques for the Design of Multilayer Filters*. A. Hilger, 1981.

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

[11] H. A. MacLeod and H. A. Macleod. *Thin-Film Optical Filters*. CRC Press, 2010.

[12] S. Martin, J. Rivory, and M. Schoenauer. Synthesis of optical multilayer systems using genetic algorithms. *Applied Optics*, 34, 1995.

[13] W. Masson, P. Ranchod, and G. Konidaris. Reinforcement learning with parameterized actions. *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, page 1934–1940, 2016.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *NIPS Deep Learning Workshop*, 2013.

[15] W. Paszkowicz. Genetic algorithms, a nature-inspired tool: A survey of applications in materials science and related fields: Part ii. *Materials and Manufacturing Processes*, 28, 2013.

[16] X. B. Peng, G. Berseth, and M. V. de Panne. Terrain-adaptive locomotion skills using deep reinforcement learning. *ACM Transactions on Graphics*, 2016.

[17] J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria, B. G. DeLacy, J. D. Joannopoulos, M. Tegmark, and M. Soljačić1. Nanophotonic particle simulation and inverse design using artificial neural networks. *Sci Adv*, 2018.

[18] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *In: Proceedings of the International Conference on Learning Representations*, 2016.

[19] B. T. Sullivan and J. A. Dobrowolski. Implementation of a numerical needle method for thin-film design. *Applied Optics*, 35, 1996.

[20] Sutton and Barto. *Introduction to reinforcement learning*. Cambridge: MIT Press, 1998.

[21] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, 2016.

[22] C. Watkins. Learning from delayed rewards. *PhD thesis*, 1989.

[23] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 1962.

[24] J. Xiong, Z. Y. Qing Wang, P. Sun, Y. Z. Lei Han, H. Fu, T. Zhang, and H. L. Ji Liu. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space. *arXiv preprint*, 2018.

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

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 indexes. Basically, a DBR is determined by two thickness values $t_1, t_2 \in \mathbb{R}^+$ and real refractive indexes $n_1, n_2 \in \mathbb{R}^+$, 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 [11]. 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

$$\Delta\lambda = \frac{4}{\pi} \cdot \lambda_0 \cdot \text{arcsin} \left| \frac{n_2 - n_1}{n_2 + n_1} \right|$$

$$\lambda_0 + \Delta\lambda = 550\,nm$$

can be solved yielding $\lambda_0 = 424.59\,nm$. 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 8-layer DBR of $473.88\,nm$ total thickness, we repeatedly stack these two layers four times.

B Impact of reward transformation and Q-value reliability

Following the optimal policy, which leads to an optimal optical system, 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

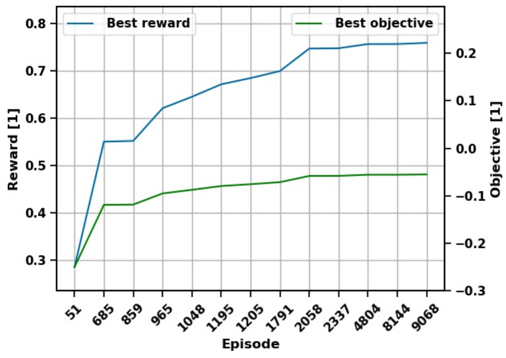


Figure 4: The best obtained objective (1) and reward over episode of its achievement for $\alpha = 18.42$ reveals the higher discriminability of designs during training. Axis limits are chosen such that the absolute length of the axes of reward and error coincide.

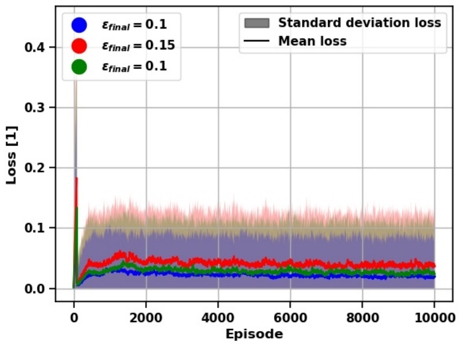


Figure 5: Illustration of the standard deviation of and mean value of the computed loss over episode. We investigated different configurations of ϵ_{final}. Note that we omitted the loss-weighted sampling of mini-batches in one case (green).

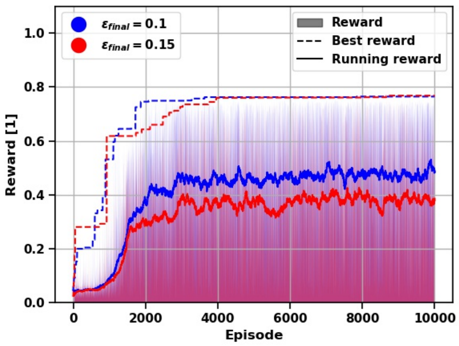


Figure 6: Illustration of obtained reward (filled area) and running reward (solid line) over episode. The best obtained reward is indicated by dashed lines for two configurations of ϵ_{final}.

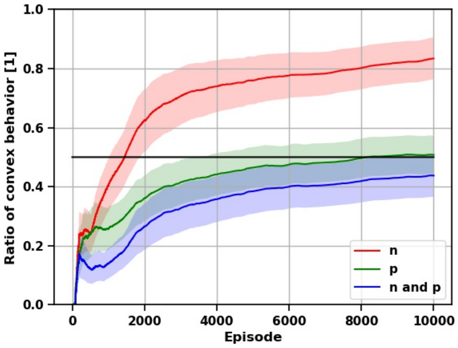


Figure 7: Step ratio of convex behavior of Q-value approximation in terms of refractive index (n), optical path length (p), and corresponding coincidence (n and p) over episode, respectively.

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 with 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 $1,000$ randomly drawn optical systems. The impact of this transformation regarding task 2 is illustrated in figure 4.

It is often not discussed that the approximation of Q-values can be monitored during policy optimization. In figure 5, 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 $5,000$ 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: 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.

C Learning dynamics

In addition to the loss, we also tracked l-step returns and eventually achieved rewards for each episode. Figure 6 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 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 $i \leq L$ 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 7 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 7. The running mean and standard deviation of ratios were computed based on Welford's online algorithm [23]. Although the optical path length intuitively gives a more encompassing optical information about a parameterized action, the ratios of convex behavior based on refractive indexes (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 indexes 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.