Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Joint Design of Adaptive Modulation and Precoding for Physical Layer Security in Visible Light Communications using Reinforcement Learning

DUC M. T. HOANG[1], THANH V. PHAM[2], (Member, IEEE), ANH T. PHAM [3], (Senior, IEEE) and ,CHUYEN T. NGUYEN[1]

[1]School of Electrical and Electronic Engineering of Hanoi University of Science and Technology, Hanoi, Vietnam (email: minhduc.hoangtrong@gmail.com, chuyen.nguyenthanh@hust.edu.vn)
[2]Department of Mathematical and Systems Engineering, Shizuoka University, Shizuoka, Japan (e-mail: pham.van.thanh@shizuoka.ac.jp)
[3]Department of Computer Science and Engineering, The University of Aizu, Fukushima, Japan (e-mail: pham@u-aizu.ac.jp)

Corresponding author: Thanh V. Pham (e-mail: pham.van.thanh@shizuoka.ac.jp).

Part of this paper has been presented at the 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), Florence, Italy [1]. This research was supported by the Japan Society for Promotion of Science (JSPS) KAKENHI under Grant 23K13333 and the Tokai Foundation for Industrial Technology Promotion.

ABSTRACT There has been an increasing interest in physical layer security (PLS), which, compared with conventional cryptography, offers a unique approach to guaranteeing information confidentiality against eavesdroppers. In this paper, we study a joint design of adaptive M-ary pulse amplitude modulation (PAM) and precoding, which aims to optimize wiretap visible-light channels' secrecy capacity and bit error rate (BER) performances. The proposed design is motivated by higher-order modulation, which results in better secrecy capacity at the expense of a higher BER. On the other hand, a proper precoding design, which can manipulate the received signal quality at the legitimate user and the eavesdropper, can also enhance secrecy performance and influence the BER. A reward function that considers the secrecy capacity and the BERs of the legitimate user's (Bob) and the eavesdropper's (Eve) channels is introduced and maximized. Due to the non-linearity and complexity of the reward function, it is challenging to solve the optical design using classical optimization techniques. Therefore, reinforcement learning-based designs using Q-learning and Deep Q-learning are proposed to maximize the reward function. Simulation results verify that compared with the baseline designs, the proposed joint designs achieve better reward values while maintaining the BER of Bob's channel (Eve's channel) well below (above) the pre-FEC (forward error correction) BER threshold.

INDEX TERMS Visible light communications, adaptive modulation, precoding, physical layer security, reinforcement learning.

I. INTRODUCTION

A. BACKGROUND

OVER the past decade, visible light communications (VLC) have emerged as a promising wireless technology for high-speed broadband communications, addressing the problem of spectrum congestion in radio frequency (RF) by shifting communications to the visible light spectrum [2]–[4]. It offers several distinct advantages, such as lower implementation costs due to its license-free spectrum and the use of existing light infrastructure, being able to operate in areas where RF signals are prohibited, and seamless integration into heterogeneous wireless networks without causing or experiencing RF interference [5]. Due to these characteristics, VLC has found applications in several areas, including indoor positioning, vehicular networks, optical camera communication, and mobile health monitoring [6], and has been integrated with RF communications [7].

Despite these potentials, VLC is vulnerable to eavesdropping by malicious users, who can exploit the broadcast nature of visible light signals to intercept the communication to a legitimate user [8]. Conventional security measures to prevent eavesdropping by using key-based cryptography. The current state-of-the-art cryptographic algorithms offer computational security and prove effective against attackers with classical computing power. However, there is an increasing concern about the security of conventional cryptography due to the rapid development of quantum computers, which can, theoretically, have an unlimited computational capability. This potential threat has spurred great interest in developing novel security approaches.

In this regard, information-theoretic security, such as physical layer security (PLS), has been the subject of active research during the last two decades. Based on the notion of information theory, PLS can protect the confidentiality of information transmitted to a legitimate user (Bob) under the presence of an eavesdropper (Eve), who is assumed to have unbounded computational power [9], [10]. Traditionally, the secrecy capacity, which defines the maximum rate the information is sent to Bob at which Eve cannot decode any information from its overheard signal, is primarily used as the performance metric of PLS. Research on PLS then focuses on characterizing the secrecy capacities of different wireless systems and enhancement methods, notably precoding and artificial noise (AN).

B. RELATED WORKS

While PLS has been extensively studied for RF systems, its adoption to VLC has received considerable attention over the past few years or so [11]. Unlike the RF counterpart, VLC signals are non-negative and subject to an amplitude limit due to the nonlinear characteristic of LEDs and to comply with the peak optical power requirement (for eye safety regulations). This signal amplitude constraint renders a derivation of closed-form expressions for the channel secrecy capacity infeasible. It has been shown in [12] that, for single-input single-output (SISO) wiretap VLC channels, the secrecy capacity-achieving distribution is discrete with a finite number of mass points. Due to this discreteness nature, a tractable closed-form expression for the secrecy capacity is yet available. Subsequent works, therefore, investigated several lower and upper bounds on the secrecy capacity [13]–[15].

Practical VLC systems for indoor settings often require the deployment of multiple LED luminaires to provide sufficient illumination, which results in the multiple-input single-output (MISO) configuration. Multiple LED transmitters enable transmit diversity, which can be exploited in the form of precoding and AN. In this regard, there have been several works investigating the design of different precoding [16]–[21] and AN schemes [22]–[25] for improving the secrecy performance of MISO VLC systems. Since VLC signals are subject to amplitude constraints that render deriving simple closed-form expressions for the secrecy capacity impossible, simple lower and/or upper bounds were extensively used to facilitate solving the optimal design.

While previous works provided valuable insights into the secrecy performance limits and the structures of optimal precoder and AN, they did not consider two practical issues. Firstly, the secrecy capacity bounds (for example, [13]–[15]) were derived without consideration of the signal modulation. Although the derived bounds are helpful in understanding the theoretical limit of the system performance, they may not provide accurate assessments in the case of practical systems where constraints on modulation (and hardware) must be considered. Secondly, the issue of communication reliability was also ignored in previous studies. A communication channel is said to be reliable if the bit error rate (BER) after forward error correction (FEC) decoding can be made sufficiently small (e.g., below 10^{-12} or so). Assuming hard decision (HD)-FEC, a certain BER must be attained before the FEC decoding to achieve this requirement. This is known as the pre-FEC BER limit[1].

In the context of wiretap channels, in addition to the communication reliability of Bob's channel, a design that simultaneously renders Eve's channel unreliable is of interest as it can further increase communication secrecy. To the best of our knowledge, only the study in [27] considered the issue of channel reliability in precoding design for wiretap VLC channels. This study, nonetheless, leaves two opening issues for further improvement. Firstly, it is the use of a modulation-independent lower bound on the secrecy capacity for the precoding design proposed in [28]. Secondly, the BER of Bob's channel was approximately calculated assuming a fixed modulation order (i.e., 4-PAM). Additionally, an unreliability constraint on Eve's channel was also ignored.

C. CONTRIBUTIONS

It is well-known that a higher secrecy rate can be achieved using higher-order modulation at the expense of a degraded BER performance [29]. Moreover, as mentioned in the above section, the secrecy performance can also be improved by utilizing precoding. Motivated by these observations, this work proposes a joint design of adaptive modulation and precoding that considers the secrecy capacity and communication reliability of Bob and Eve's channels. Specifically, the proposed design aims to maximize the secrecy capacity of M-ary pulse amplitude modulation (M-PAM) while guaranteeing the communication reliability of Bob's channel and the communication unreliability of Eves's channel, respectively. For this purpose, a reward function that captures a tradeoff among the secrecy capacity of M-PAM, the BERs of Bob's, and Eve's channels is introduced and maximized, given the constraint on the amplitude of VLC signals. Such an optimization problem, however, is difficult (if not impossible) to solve using classical optimization techniques due to the unavailability of a closed-form expression for the PAM-constrained secrecy capacity and the complex nonlinear ex-

[1] It should be noted that the pre-FEC BER limit is a valuable predictor for the BER after FEC decoding in the case of HD-FEC. However, as verified in [26], the generalized mutual information (GMI) is a better predictor when soft decision (SD)-FEC is employed.

pressions of the BERs. In this paper, therefore, we tackle the design problem using reinforcement learning approaches.

The main contributions of this work are as follows:

- We proposed a joint design of adaptive M-PAM modulation and precoding for MISO VLC channels. The proposed design aims to enhance the secrecy capacity while ensuring the reliability of Bob's channel and, at the same time, degrade the reliability of Eve's channel.
- Reinforcement learning approaches based on Q-learning and Deep Q-learning are employed to jointly optimize the modulation order and the transmit precoder. Specifically, to this end, a reward function that captures the secrecy capacity, the BERs of Bob's and Eve's channels, is introduced.
- The exact secrecy capacity expression of M-PAM is derived in integral form, which is computationally expensive, especially in the case of high modulation order. This, in turn, leads to an increased training time for the proposed learning approaches. An approximate expression for the secrecy capacity is presented to reduce the training time. Extensive comparisons of the accuracy and computational time between the exact and approximate expressions are then performed.
- Comprehensive simulations are carried out to determine the appropriate weight factors of the reward function and to illustrate the superiority of reinforcement learning-based joint designs over the baseline schemes.

D. ORGANIZATION

The paper is structured as follows. In Section II, the system model, including the channel and signal models, is given. Section III presents the exact and approximate expressions for the secrecy capacity and bit error rate (BER) of M-PAM modulation. The proposed Q-learning and deep Q-learning-based reinforcement learning approaches for the design of joint adaptive modulation and precoding are described in Sections IV and V, respectively. Simulation results and comparisons with baseline schemes are discussed in Section VI. Finally, Section VII concludes the paper.

Notation: We extensively use the following notation in the paper. Column vectors are represented by lowercase bold letters (e.g., $\mathbf{x}$), and the transpose of $\mathbf{x}$ is denoted as $\mathbf{x}^T$. In addition, $\mathbf{1}_N$ is the all-ones vector of size N, and $\mathbb{R}$ is the set of real numbers. Finally, $I(\cdot;\cdot)$, $h(\cdot)$, and $p(\cdot)$ denote the mutual information, differential entropy, and probability, respectively.

II. SYSTEM MODEL

A. CHANNEL MODEL

A VLC system, which is composed of N LED luminaires acting as the transmitter (Alice), a legitimate user (Bob), and an eavesdropper (Eve), is examined in this study. We assume that Bob and Eve have a single photodiode (PD) receiver. In some scenarios, such as multi-user systems where the confidentiality of the information intended for Bob must be kept, all other active users should be considered as Eves. Their channel state information (CSI), therefore, can be known at the transmitter. Moreover, from the analysis point of view, the assumption of known Eve's CSI gives the upper bound performance. Hence, we assume in this study that the CSI of both Bob's and Eve's channels are known [16]–[25][2]. The overall system configuration is shown in Fig. 1.

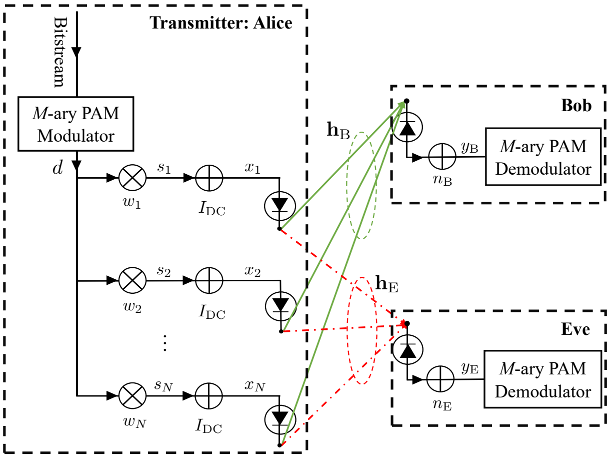


FIGURE 1: System model.

Due to signal reflections off walls and ceilings, VLC channels in indoor environments usually consist of two components: the line-of-sight (LoS) and nonline-of-sight (NLoS). Nonetheless, the previous study in [30] showed that the LoS propagation contributes to, on average, more than 95% of the total optical power received by the receiver. Hence, to simplify the analysis without considerably affecting the evaluation accuracy, only the LoS channel is considered in this study.

Let us denote $\mathbf{h}_{\text{R}} = \begin{bmatrix} h_{\text{R}}^1 & h_{\text{R}}^2 & \cdots & h_{\text{R}}^N \end{bmatrix}^T \in \mathbb{R}_{\geq 0}^N$ as the channel vector of user R[3] where the n-th element h_{R}^n is the LoS channel gain between the n-th luminaire and the receiver. Note that most LED luminaires have the Lambertian beam distribution, whose emission intensity concerning an angle of irradiance ϕ is given by

$$L(\phi) = \frac{l+1}{2\pi}\cos^l(\phi), \tag{1}$$

where $l = \frac{\ln(2)}{\ln(\Theta_{0.5})}$ with $\Theta_{0.5}$ being the LED's semi-angle for half illuminance is the Lambertian emission order. According to [30], the channel gain h_{R}^n is then

$$h_{\text{R}}^n = \begin{cases} \frac{A_r}{d_n^2} L(\phi) T_s(\psi) g(\psi)\cos(\psi), & 0 \leq \psi \leq \Psi, \\ 0, & \psi > \Psi, \end{cases} \tag{2}$$

where A_r is the active area of the PD, d_n is the distance between the luminaire and the PD, ψ is the angle of incidence, and Ψ denotes the optical field of view (FoV) of the

[2] When the presence of Eve is unknown to the transmitter, an average channel gain of Eve can be calculated and used (for example, see [23, Eq. (12)]). For conciseness, we omit this case in the paper.

[3] The subscript 'R' is used to refer to user R in general. However, when needed, it is replaced by 'B' and 'E' to explicitly denote Bob and Eve, respectively.

PD. Furthermore, the gain of the optical filter is indicated by $T_s(\psi)$, and $g(\psi)$ is the gain of the optical concentrator, which is given by

$$g(\psi)=\begin{cases}\frac{\kappa^2}{\sin^2(\Psi)}, & 0\leq\psi\leq\Psi,\\ 0, & \psi>\Psi,\end{cases} \tag{3}$$

where κ is the refractive index of the optical concentrator.

B. SIGNAL MODEL

Due to its simplicity and straightforward implementation, M-PAM is widely used in VLC systems. On the transmitting side, as shown in Fig. 1, the input bits are fed to an M-PAM modulator to generate bipolar PAM symbols, denoted as d. At the n-th luminaire, the PAM symbol is first precoded using a precoder w_n. Since the input signal of the LEDs must be nonnegative, a DC-bias current I_{DC} should be added to the precoded signal. The addition of DC bias is also to adjust the illumination level of the LEDs (Note that for VLC systems, illumination is still the primary purpose that the performance of the communications aspect should be investigated under proper illumination settings). The drive current of the LEDs is thus written by

$$x_n=w_nd+I_{\text{DC}}. \tag{4}$$

It should be noted that LEDs exhibit a specific linear range over which the emitted optical power is proportional to the amplitude of the drive current. To maintain this linear conversion (for the sake of proper LED operation and energy efficiency), x_n should be constrained within some range $I_{\text{DC}}\pm\alpha I_{\text{DC}}$ where $\alpha\in[0,\ 1]$ denotes the modulation index [28], [31]. To satisfy these constraint on x_n, one has $|w_n|\leq 1$ and $|d|\leq\alpha I_{\text{DC}}$.

If we let $\mathbf{w}=\begin{bmatrix}w_1 & w_2 & \cdots & w_N\end{bmatrix}^T\in\mathbb{R}^N$ be the transmit precoder, the transmitted signal vector is given by

$$\mathbf{x}=\mathbf{w}d+\mathbf{1}_NI_{\text{DC}}, \tag{5}$$

where $\mathbf{x}=\begin{bmatrix}x_1 & x_2 & \cdots & x_N\end{bmatrix}^T\in\mathbb{R}^N$. The emitted optical power vector is then given by

$$\mathbf{p}_t=\eta\mathbf{x}=\eta(\mathbf{w}d+\mathbf{1}_NI_{\text{DC}}), \tag{6}$$

where η is the electrical-to-optical conversion efficiency of the LEDs. Denoting γ as the PD responsivity, the electrical signal at the output of the PD is written by

$$\begin{aligned}y_{\text{R}}&=\gamma\mathbf{h}_{\text{R}}^T\mathbf{p}_t+n_{\text{R}}\\&=\gamma\eta\mathbf{h}_{\text{R}}^T\left(\mathbf{w}d+\mathbf{1}_NI_{\text{DC}}\right)+n_{\text{R}}.\end{aligned} \tag{7}$$

Here, n_{R} is the receiver noise, which includes thermal and shot noises. It is generally valid to assume that n_{R} is a zero-mean additive white Gaussian noise with variance σ_{R}^2 [19], [23]. For signal modulation, the DC-term $\mathbf{h}_{\text{R}}^T\mathbf{1}_NI_{\text{DC}}$, which contains no information, is filtered out, resulting in

$$\overline{y}_{\text{R}}=\gamma\eta\mathbf{h}_{\text{R}}^T\mathbf{w}d+n_{\text{R}}. \tag{8}$$

Since each element of $\mathbf{w}$ must be bounded between $[-1,\ 1]$, the constraint on the precoder $\mathbf{w}$ can be represented as

$$\|\mathbf{w}\|_\infty\leq 1, \tag{9}$$

where $\|\cdot\|_\infty$ is the infinity norm. On the contrary, the transmitted signals in RF systems are often subject to average power constraints, which is represented by the Euclidean norm of the precoding vector [32].

III. SECRECY CAPACITY AND BIT ERROR RATE

A. SECRECY CAPACITY

The secrecy capacity C_s of the considered wiretap channel is given as the difference between the capacities of Bob's channel (denoted as C_{B}) and Eve's channel (denoted as C_{E}) [10], i.e.,

$$C_s=\left[C_{\text{B}}-C_{\text{E}}\right]^+, \tag{10}$$

where $[x]^+\triangleq\max(x,0)$. Based on (8), these channel capacities are defined as the mutual information between the channel input d and the channel output $\overline{y}_{\text{R}}$

$$C_{\text{R}}=I(d;\overline{y}_{\text{R}}), \tag{11}$$

where we have

$$I(d;\overline{y}_{\text{R}})=h(\overline{y}_{\text{R}})-h(\overline{y}_{\text{R}}|d)=h(\overline{y}_{\text{R}})-h(n_{\text{B}}), \tag{12}$$

where differential entropies of $\overline{y}_{\text{R}}$ and n_{R} are given by

$$h(\overline{y}_{\text{R}})=-\int_{-\infty}^{+\infty}p(\overline{y}_{\text{R}})\text{log}_2p(\overline{y}_{\text{R}})\text{d}\overline{y}_{\text{R}}, \tag{13}$$

$$h(n_{\text{R}})=\frac{1}{2}\text{log}_2\left(\pi e\sigma_{\text{R}}^2\right), \tag{14}$$

respectively. Let $\{d_i\}$ with $i=0,1,\cdots,M-1$ be the set of PAM constellation symbols and assume that all symbols are transmitted with equal probabilities (i.e., $p(d_i)=\frac{1}{M}$, $\forall i=0,1,\cdots,M-1$). Since n_{R} is Gaussian with zero-mean and variance σ_{R}^2, $p(\overline{y}_{\text{R}})$ is given by

$$\begin{aligned}p(\overline{y}_{\text{R}})&=\sum_{i=0}^{M-1}p(\overline{y}_{\text{R}}|d=d_i)p(d_i)\\&=\frac{1}{M}\sum_{i=0}^{M-1}\frac{1}{\sqrt{2\pi}\sigma_{\text{R}}}\exp\left(-\frac{\left|\overline{y}_{\text{R}}-\gamma\eta\mathbf{h}_{\text{R}}^T\mathbf{w}d_i\right|^2}{2\sigma_{\text{R}}^2}\right).\end{aligned} \tag{15}$$

From the above analysis, it is seen that the evaluation of the secrecy capacity in (10) involves two integral operations in (13), which are generally time-consuming, especially when the integrand is complicated (as can be seen from (15), the complexity of the integrand is directly proportional to modulation order. It is thus expected that the execution time increases with the modulation order, which we later numerically verify). As a result, using the exact expression of the secrecy capacity may incur a long training time, as reinforcement learning algorithms usually require a large number of iterations until convergence. To reduce the training time, we present an approximation, which does not require integration, to the secrecy capacity. To this end, (15) is rewritten by

$$p(\overline{y}_{\mathrm{R}}) = \frac{1}{M}\sum_{i=0}^{M-1}\mathcal{N}(\overline{y}_{\mathrm{R}}, \mu_i, \sigma_{\mathrm{R}}), \tag{16}$$

where $\mu_i = \gamma\eta\mathbf{h}_{\mathrm{R}}^T\mathbf{w}d_i$ and $\mathcal{N}(x, \mu, \sigma)$ denotes the probability density function (PDF) of the Gaussian variable x with mean μ and standard deviation σ.

The Taylor expansion of the logarithm component in (13) gives

$$\log_2 p(\overline{y}_{\mathrm{R}}) = \sum_{k=0}^{\infty}\frac{1}{k!}\left(\overline{y}_{\mathrm{R}} - \mu_i\right)^k \frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}, \tag{17}$$

with $\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}$ being the k-th order derivative of $\log_2 p(\overline{y}_{\mathrm{R}})$ with respect to $\overline{y}_{\mathrm{R}}$ and $\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}$ denoting the value of $\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}$ at $\overline{y}_{\mathrm{R}} = \mu_i$. Replacing (16) and (17) into (13) yields

$$\begin{aligned} h(\overline{y}_{\mathrm{R}}) &= \frac{-1}{M}\sum_{i=0}^{M-1}\int_{-\infty}^{+\infty}\mathcal{N}(\overline{y}_{\mathrm{R}}, \mu_i, \sigma_R) \\ &\times\left(\sum_{k=0}^{\infty}\frac{1}{k!}\left(\overline{y}_{\mathrm{R}} - \mu_i\right)^k \frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}\right)\mathrm{d}\overline{y}_{\mathrm{R}} \\ &= \frac{-1}{M}\sum_{i=0}^{M-1}\sum_{k=0}^{\infty}\frac{\rho_{i,k}}{k!}\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}, \end{aligned} \tag{18}$$

where $\rho_{i,k} = \int_{-\infty}^{\infty}(\overline{y}_{\mathrm{R}} - \mu_i)^k\mathcal{N}(\overline{y}_{\mathrm{R}}, \mu_i, \sigma_{\mathrm{R}})\mathrm{d}\overline{y}_{\mathrm{R}}$ is the Gaussian k-th central moment, which is given by

$$\rho_{i,k} = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \sigma_{\mathrm{R}}^k(k-1)!! & \text{if } n \text{ is even,} \end{cases} \tag{19}$$

with $n!!$ being the double factorial [33]. Note that the expression in (18) is an infinite series, which can be truncated to get an approximation. Through an extensive evaluation, it is found that using the first three terms offers satisfactory accuracy with low computational complexity. An approximation to (18) is thus

$$h(\overline{y}_{\mathrm{R}}) \approx \frac{-1}{M}\sum_{i=0}^{M-1}\sum_{k=0}^{2}\frac{\rho_{i,k}}{k!}\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}. \tag{20}$$

Following (19), we have $\rho_{i,0} = 1$, $\rho_{i,1} = 0$, and $\rho_{i,2} = \sigma_{\mathrm{R}}^2$. The above expression is thus simplified to

$$h(\overline{y}_{\mathrm{R}}) \approx \frac{-1}{M}\sum_{i=0}^{M-1}\left(\log_2 p(\mu_i) + \frac{\sigma_{\mathrm{R}}^2}{2}\frac{d^k \log_2 p(\overline{y}_{\mathrm{R}})}{d\overline{y}_{\mathrm{R}}{}^k}\,|_{\overline{y}_{\mathrm{R}}=\mu_i}\right). \tag{21}$$

The accuracy of the presented approximation is numerically demonstrated in Fig. 2 for different modulation orders and noise variances over a wide range of the signal-to-noise ratio (SNR). Here, the SNR is defined as $\mathrm{SNR} = 10\log_{10}\frac{\left(\gamma\eta\mathbf{h}_{\mathrm{R}}^T\mathbf{w}E_s\right)^2}{\sigma_{\mathrm{R}}^2}$ (dB) with $E_s = \mathbb{E}[d]$ being the average symbol energy. It is observed that the approximate capacity provides a satisfactory agreement with the exact value over wide ranges of the SNR, modulation order, and noise variances.

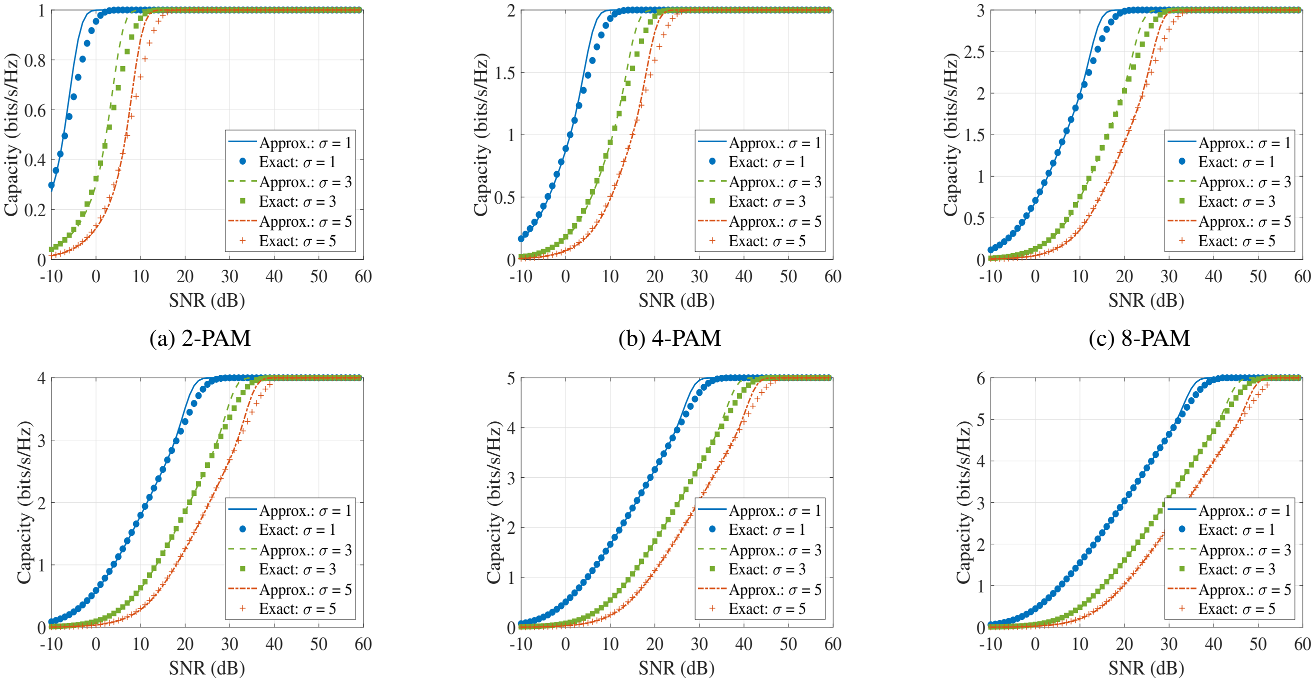
Comparisons regarding the computational complexities of the approximate and exact expressions are also tabulated in Table 1 where the execution time is averaged over 100 evaluations with SNR $=$ 20 dB and $\sigma = 3$. Calculations are performed using Python on Google Colaboratory on a Windows 11 desktop computer with Intel® Xeon® processor 2.30GHz and 12GB RAM. It is clearly shown that the execution time of the approximate expression is significantly faster than that of the exact one, especially when the modulation order M is small. Specifically, in the case of 2-PAM, the expectation time of the approximate expression is nearly 15 times faster. It is also noticed that as the modulation order increases, the difference in the execution time between the approximate and exact expressions becomes less yet still significant. For example, at 64-PAM, the execution time of the approximation is about 1.2 times faster. However, it is worth noting that even a small improvement in the calculation time of the secrecy capacity can significantly reduce the overall training time of the proposed reinforcement learning approaches since the learning process generally takes a few thousand iterations until convergence.

B. BIT ERROR RATE

Assuming the Gray coding, the BER of M-PAM is given by [34]

TABLE 1: Execution time (in seconds) of the approximate and exact capacity expressions.

Modulation order M **Execution time, (seconds)**	2	4	8	16	32	64
Approximate expression in (20)	0.06	0.26	0.53	1.79	6.70	24.87
Exact expression in (10)	0.89	0.95	3.07	4.91	13.91	30.32



(a) 2-PAM (b) 4-PAM (c) 8-PAM

(d) 16-PAM (e) 32-PAM (f) 64-PAM

FIGURE 2: Approximation of the capacity with different modulation orders and noise variances.

$$p_{\mathrm{e,R}} = \sum_{k=1}^{\log_2 M}\sum_{i=0}^{M(1-2^{-k})-1}(-1)^{\left\lfloor \frac{i2^{k-1}}{M}\right\rfloor}\left(2^{k-1}-\left\lfloor \frac{i2^{k-1}}{M}+\frac{1}{2}\right\rfloor\right) \times \frac{1}{M\log_2 M}\mathrm{erfc}\left((2i+1)\sqrt{\frac{3\left(\gamma\eta\mathbf{h}_{\mathrm{R}}^{T}\mathbf{w}/\sigma_{\mathrm{R}}\right)^{2}E_s}{M^2-1}}\right), \tag{22}$$

where $\lfloor\cdot\rfloor$ denotes the floor function and $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$ is the complimentary error function.

IV. PROPOSED JOINT DESIGN

This study aims at a joint design of adaptive modulation and precoding that maximizes secrecy capacity while guaranteeing the reliability of Bob's channel and rendering Eve's channel unreliable. It has been shown that increasing the modulation order improves the secrecy capacity [29]. Nonetheless, higher modulation also makes symbols more susceptible to noise (due to reduced distance between symbols), resulting in a higher BER and, hence, less reliability. On the other hand, previous studies have also shown that dedicated precoding designs can improve the secrecy capacity. Yet, they have not considered the issue of channel reliability mainly because BER expressions are often nonlinear, non-convex, and highly complex (for example, the one in (22) for the case of M-PAM). We note here that, due to the contrasting impact of modulation order on the secrecy capacity and the channel reliability, as well as the non-convexity and high complexity of the BER expressions, the proposed joint design may not be explicitly solved using classical optimization techniques. The difficulty of a rigorous mathematical solution motivates us to investigate the use of machine learning-based techniques. In particular, reinforcement learning (RL) approaches based on Q-learning and Deep Q-learning are considered in this study.

In this work, we only focus on the offline training phase as in line with most RL-based wireless system designs, for example, [35]–[37]. Specifically, the agent is trained using a dataset generated in a simulated environment. The dataset used is generated by randomizing the positions of Bob and Eve, allowing the agent to learn the optimal transmission policy in every Bob and Eve location. The optimal policy is then transferred to the practical system in the online deployment phase. At this phase, the agent can choose the optimal action directly after observing the state, enabling Alice to choose modulation order and precoding in real time. Furthermore, Alice continuously updates and fine-tunes its policy with the CSI received from the feedback channel for Bob and Eve in each specific configuration. This iterative process enables our approach to adapt to changing channel conditions. Further details on these algorithms are provided in the next sections.

A. Q-LEARNING-BASED JOINT DESIGN

Q-learning is a popular model-free reinforcement learning algorithm that aims to achieve the optimal action policy for an agent interacting with an environment through maximizing a reward function [38], [39]. In the proposed joint design context, the agent, environment, state, action, and reward function are specified as follows.

- **Agent**: The central processing unit (CPU) at the transmitter that is responsible for optimizing the modulation order M and precoder $\mathbf{w}$.
- **Environment**: The entire system except for the CPU, including, for example, the LED luminaires' optical power, positions of Bob and Eve, and BERs of Bob's and Eves' channels.
- **State**: A subset of the system's parameters that are the input for the agent to select the appropriate action. Specifically for the proposed design, the state $\mathbf{s}$ comprises the BERs of Bob's and Eve's channels at the previous state and the CSI of Bob's and Eve's channels at the current state. Let Λ be the set of all possible states.
- **Action**: An action, denoted by $\mathbf{a}$, is the selection of the modulation order M and the precoder $\mathbf{w}$. Moreover, we define $\mathbb{A}$ as the set of all possible actions. For the modulation order, $M = 2,\ 4,\ 8,\ 16,\ 32$ and 64 are considered. For the precoder, according to the constraint in (9), each element w_i $(i = 1, 2, \cdots, N)$ of $\mathbf{w}$ is constrained between $[-1,\ 1]$. As Q-learning requires a finite-size action space, w_i is quantized into $2T_s + 1$ equally spaced discrete values, that is, $w_i \in \left\{\frac{t}{T_s} \,\middle|\, -T_s \leq t \leq T_s,\ T_s \text{ and } t \in \mathbb{Z}\right\}$. Note that this action quantization sets a trade-off between the performance and training time, as increasing the size of the action space by increasing T_s leads to a better performance at the expense of longer training time.
- **Reward function**: As the reliability of a communication channel is defined by its BER performance, we consider the following reward function, which captures the secrecy capacity, the BERs of Bob's and Eve's channel
$$u = C_s - \delta p_{e,\mathrm{B}} + \zeta p_{e,\mathrm{E}}. \tag{23}$$
Here, δ and ζ are chosen coefficients that influence the contribution of BERs of Bob and Eve's channels on the reward value, respectively. The selection of these parameters is numerically discussed in Section V. The motivation for the definition in (23) is that the maximization of u is equivalent to maximizing C_s and $p_{e,\mathrm{E}}$ while minimizing $p_{e,\mathrm{B}}$. This is indeed in accordance with the objective of the proposed design.

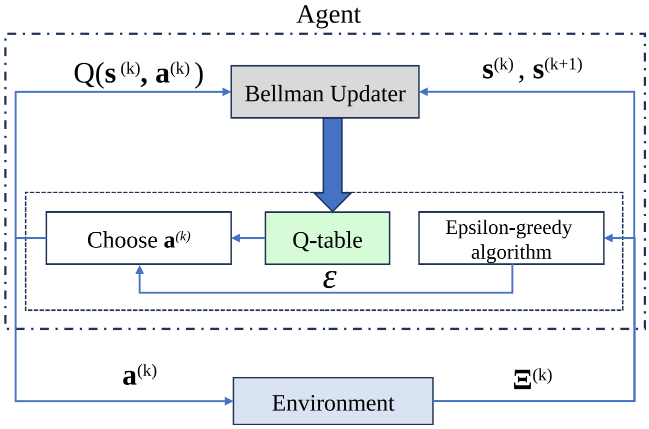


FIGURE 3: Proposed Q-learning-based joint design.

The overall schematic diagram of the proposed Q-learning-based joint design is illustrated in Fig. 3, where a Q-table was used to store the value of the state-action pair. At the time slot k of the training, the agent selects an action $\mathbf{a}^{(k)} = \left[M^{(k)}, \mathbf{w}^{(k)}\right]$ based on the current state $\mathbf{s}^{(k)}$, which consists of the BERs of Bob's and Eve's channel, the secrecy capacity at the previous time slot and the channel vectors of Bob and Eve at the current time slot, i.e., $\mathbf{s}^{(k)} = \left[p_{e,\mathrm{B}}^{(k-1)}, p_{e,\mathrm{E}}^{(k-1)}, C_s^{(k-1)}, \mathbf{h}_{\mathrm{B}}^{(k)}, \mathbf{h}_{\mathrm{E}}^{(k)}\right]$. After taking the action, the agent observes the environment to gather the current experience $\Xi^{(k)} = \left[\mathbf{s}^{(k)}, \mathbf{a}^{(k)}, u^{(k)}, \mathbf{s}^{(k+1)}\right]$, which consists of the current state, action, reward, and the next state. Using this current experience, the agent constructs the Q-value table and updates it using the Bellman equation [40]. Specifically, the Q-value at the k-th time slot is given by

$$Q\left(\mathbf{s}^{(k)}, \mathbf{a}^{(k)}\right) = (1-\lambda) Q\left(\mathbf{s}^{(k)}, \mathbf{a}^{(k)}\right) + \lambda\left(u^{(k)} + \beta \max_{\mathbf{a}} Q\left(\mathbf{s}^{(k+1)}, \mathbf{a}^{(k+1)}\right)\right). \tag{24}$$

Here, $\lambda \in [0, 1]$ is the learning rate, which determines to what extent the Q-value is updated after each time slot. The learning rate sets a trade-off between the convergence speed and the stability of the learning algorithm (a low learning rate often results in a slow convergence, while a high one may cause the learning to oscillate or to fail to converge). The parameter $\beta \in [0, 1]$ is the discount factor, which determines the importance of the immediate and future rewards. A low β implies that the immediate reward is preferable, whereas a high β is chosen when the agent aims for a long-term reward.

Algorithm 1 Q-learning-based joint design

```
$Q(\mathbf{s}, \mathbf{a}) = 0,\ \forall \mathbf{s} \in \Lambda,\ a \in \mathbb{A}$.
Initialize $\mathbf{a}^{(0)} = [M^{(0)},\ \mathbf{w}^{(0)}]$, learning rate $\lambda$, and discount factor $\beta$.
for $k = 1, 2, 3, ...$ do
    Observe the environment for experience $\Xi^{(k)} = \{\mathbf{s}^{(k)}, \mathbf{a}^{(k)}, \mathbf{u}^{(k)}, \mathbf{s}^{(k+1)}\}$.
    Feed the current experience to the Bellman Updater to form the Q-table.
    Choose an action $\mathbf{a}^{(k)} = [M^{(k)},\ \mathbf{w}^{(k)}]$ using the $\epsilon$-greedy method.
    Transmit the signal with the action $\mathbf{a}^{(k)}$.
    Calculate the utility and update the Q-value according to (24).
    Calculate the current BERs $p_{\mathrm{e,B}}^{(k)}$, $p_{\mathrm{e,E}}^{(k)}$ and the secrecy capacity $C_s^{(k)}$.
end for
```

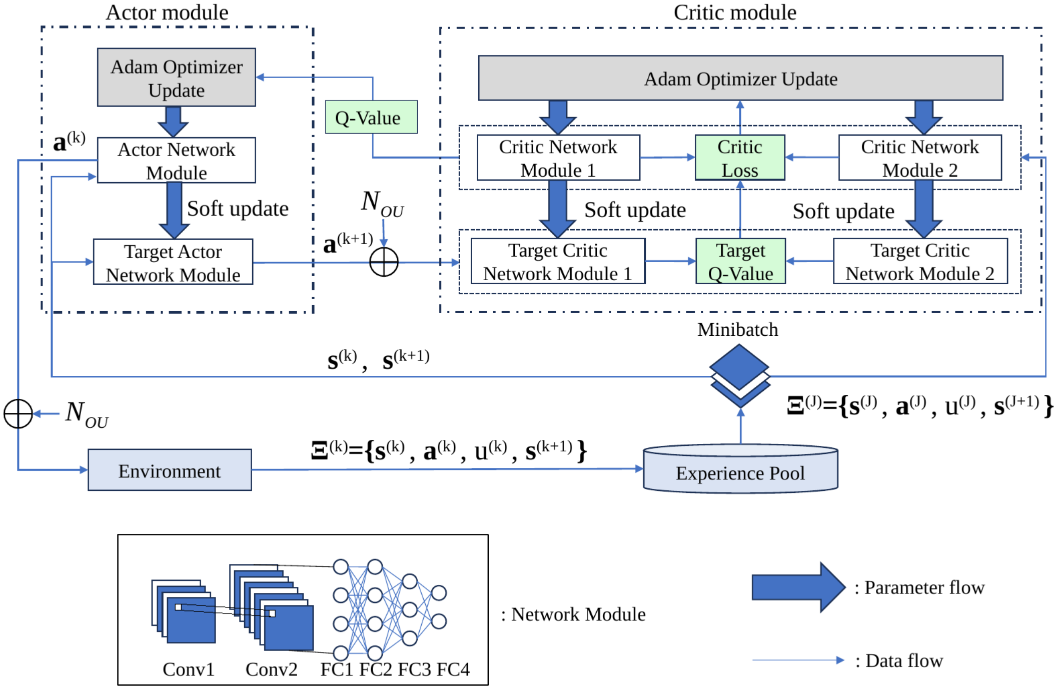


FIGURE 4: Deep reinforcement learning-based joint design.

To optimize the action selection, the agent is assumed to employ the ϵ-greedy algorithm, which aims at balancing between exploration and exploitation. Specifically, the probability of choosing the action of the k-th time slot is given by the following rule

$$P\left(\mathbf{a}^{(k)}=\mathbf{a}^{*}\right)=\begin{cases}1-\varepsilon, & \mathbf{a}^{*}=\arg\max_{\mathbf{a}'} Q\left(\mathbf{s}^{(k)},\mathbf{a}'\right)\\ \varepsilon, & \text{otherwise.}\end{cases} \tag{25}$$

Here, the algorithm takes random actions with a probability of $\epsilon \in [0,1]$ to explore the environment and avoid getting trapped in local optima. As the number of time slots increases and ϵ decreases, the algorithm increasingly favors selecting the action that maximizes the Q value with the probability of $1-\epsilon$, exploiting the highest existing Q value, and optimizing the modulation order and precoder. A summary of the Q-learning-based joint design scheme is given in **Algorithm 1**.

B. DEEP Q-LEARNING-BASED JOINT DESIGN

The proposed Q-learning-based design presented in the previous section requires a finite action space. To accomplish this requirement, the precoder $\mathbf{w}$ was quantized into discrete values, leading to a suboptimality of the action policy. Although a better action policy can be achieved by increasing the quantization resolution (i.e., $2T_s+1$), this also increases the training time.

To address this issue, we present a deep Q-learning model that capitalizes on an actor-critic network framework to tackle the continuous action space. Our approach adopts an upgraded version of the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. Specifically, TD3 was recently proposed in [41] as an improvement over the Deep Deterministic Policy Gradient (DDPG) algorithm [42] for continuous action spaces. A major advantage of TD3 compared to conventional deep reinforcement learning (DRL) algorithms is that it provides better overall performance with delayed policy updates. Instead of updating the actor policy every time step, TD3 updates once every τ time slot. Similar to its predecessor, DDPG, TD3 employs target networks to stabilize the training. These target networks, both for the actor and critic, are updated at a slower rate than the main networks, which helps reduce the variance during training.

As illustrated in Fig. 4, all of the network modules have the same architecture consisting of two convolutional neural networks (CNNs) and four full-connected (FC) layers. This configuration serves specific purposes related to the multi-dimensional nature of the problem. While a normal neural network can process data sequentially and has limitations in handling spatial data, CNNs can efficiently capture spatial hierarchies and relationships among data points, which are influenced by Bob's and Eve's location. Furthermore, combining CNNs with FC neural networks also allows a hybrid architecture that takes advantage of the strengths of both types of networks.

The framework includes two main modules. First, the actor module included an Adam optimizer [43] that updates these parameters to the actor module. The actor network module directly interfaces with the environment, allowing it to select actions and interact with the surroundings. It plays a central role in the policy-determination process, using the actor neural network function $\mu(\mathbf{s}|\theta_\mu)$ with parameter θ_μ to generate action choices based on the current state. The purpose of the target actor network is to provide a target from which the actor can learn using the function $\mu'(\mathbf{s}|\theta_{\mu'})$, but it does so with a time delay to mitigate overestimation issues commonly encountered in reinforcement learning algorithms.

The critic module relies on the output of the target actor module $\mathbf{a} = \mu'(\mathbf{s}|\theta_{\mu'})$ to assess the quality of the actions taken in the given state. It estimates the expected cumulative rewards of the Q-values associated with these actions using critic functions $Q(\mathbf{s},\mathbf{a}|\theta_{Q_1})$ and $Q(\mathbf{s},\mathbf{a}|\theta_{Q_2})$. This information is then used to guide the actor towards actions that maximize the expected cumulative rewards. During the training, the critic module employs the mean squared error loss to compare its Q-value with target Q-value which is generated by the target critic modules functions $Q'(\mathbf{s},\mathbf{a}|\theta_{Q'_1})$ and $Q'(\mathbf{s},\mathbf{a}|\theta_{Q'_2})$. These target critic networks are updated in a *soft* manner, meaning that they are adjusted toward the Q-value estimated by the critic networks, but with a slower rate of change. This soft update mechanism prevents abrupt shifts in the target critic values and promotes smoother convergence during training. This iterative process allows the critic to fine-tune its Q-value estimates, providing valuable feedback for the actor's policy optimization.

Algorithm 2 outlines the detailed data flow of the learning scheme. First, the parameters of the actor and critic networks θ_μ and $\theta_{Q_1}, \theta_{Q_2}$ are initialized and then copied to create the target networks' parameters $\theta_{\mu'}$ and $\theta_{Q'_1}, \theta_{Q'_2}$. The first action $\mathbf{a}^{(0)}$ is also randomly chosen to form the first experience $\Xi^{(0)} = \left\{\mathbf{s}^{(0)}, \mathbf{a}^{(0)}, \mathbf{u}^{(0)}, \mathbf{s}^{(1)}\right\}$ and start the loop in the algorithm. At the k-th iteration, the state $\mathbf{s}^{(k)}$ is calculated and fed to the actor network. The output $\mathbf{a}^{(k)} = \mu(\mathbf{s}^{(k)}|\theta_\mu)$ is added with the noise N_{OU} drawn from an Ornstein-Uhlenbeck (OU) noise process, which is a time-correlated noise and is shown to enable the agent better exploring the environment compared with Gaussian noise [44] [45]. The action $\mathbf{a}^{(k)}$ then affects the environment and forms the experience $\Xi^{(k)}$. The environment then stores the experience $\Xi^{(k)}$ to the experience pool, which, in our case, is a 5000-slot buffer. To update the parameters, the agent randomly extracts J batches (denoted as $\Xi^{(J)}$) from the pool and then passes them through the critic networks and target critic networks. The output of the target critic networks Q_{target} is the smaller value between two target Q-values $Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'_1}\right)$

Algorithm 2 Deep Q-learning-based joint design

```
Initialize parameters $\theta_{Q_1}$, $\theta_{Q_2}$, $\theta_{{Q'}_1}$, $\theta_{{Q'}_2}$, $\theta_\mu$, $\theta_{\mu'}$ for neural networks.
$Q(\mathbf{s},\mathbf{a}) = 0,\ \forall \mathbf{s} \in \Lambda,\ \mathbf{a} \in \mathbb{A}$.
Initialize $\mathbf{a}^{(0)} = [M^{(0)}, \mathbf{w}^{(0)}]$, learning rate $\lambda$, and discount factor $\beta$.
Initialize the experience pool
for $k = 1, 2, 3, ...$ do
    Observe the environment for experience $\Xi^{(\mathbf{k})} = \{\mathbf{s}^{(k)}, \mathbf{a}^{(k)}, u^{(k)}, \mathbf{s}^{(k+1)}\}$.
    Add the current experience to the experience pool.
    Draw random $J$ batches $\Xi^{(J)}$ from the experience pool.
    Feed $\Xi^{(J)}$ to critic network modules to get the Q-value, i.e., $\min\Big(Q\left(\mathbf{s}^{(j+1)}, \mathbf{a}^{(j+1)}|\theta_{Q_1}\right), Q\left(\mathbf{s}^{(j+1)}, \mathbf{a}^{(j+1)}|\theta_{Q_2}\right)\Big)$.
    Feed $\Xi^{(J)}$ to the target actor network module to get $\mu'\left(\mathbf{s}^{(j+1)}|\theta_{\mu'}\right)$.
    Add $\mathbf{a}^{(j+1)}$ with noise $N_{OU}$ and feed to target critic network modules to get target Q-values .
    Calculate the critic loss using Q-value and target Q-value $Q_{\text{target}}$, i.e.,
    $\min\Big(Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'_1}\right),$
        $Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'_2}\right)\Big)$.
    Use Adam optimizer to update critic network modules using (26).
    if $k$ mod $\tau == 0$ then
        Use Adam optimizer to update actor network modules using (27).
        Soft-update the target networks using (28), (29), and (30).
    end if
end for
```

TABLE 2: Deep Q-learning parameter.

Network	**Actor Network**				**Crtitic Network**			
Layer	Conv. 1	Conv. 2	FC 1	FC 4	Conv. 1	Conv. 2	FC 1	FC 4
Input size	$1 \times 42 \times 1$	$16 \times 39 \times 1$	1152	150	$1 \times 52 \times 1$	$16 \times 49 \times 1$	2208	150
Number of the filters	16	32	200	1	16	32	200	1
Output size	$16 \times 39 \times 1$	$32 \times 36 \times 1$	200	1	$16 \times 49 \times 1$	$32 \times 46 \times 1$	200	1

and $Q'\left(\mathbf{s}^{(j+1)}, \mu'(\mathbf{s}^{(j+1)}|\theta_{\mu'})|\theta_{Q'_2}\right)$. The weights θ_{Q_1}, θ_{Q_2} of critic networks are then updated to minimize the critic loss, i.e.,

$$\theta_{Q_1}, \theta_{Q_2} = \underset{\theta_{Q_1}, \theta_{Q_2}}{\arg\min} \sum_{j=1}^{J} \left(u^{(j)} + \beta Q_{\text{target}} - Q\left(\mathbf{s}^{(j)}, \mathbf{a}^{(j)}|\theta_{Q_1}\right)\right)^2 \quad (26)$$
$$+ \left(u^{(j)} + \beta Q_{\text{target}} - Q\left(\mathbf{s}^{(j)}, \mathbf{a}^{(j)}|\theta_{Q_2}\right)\right)^2,$$

where β is the discount factor. Similarly, the actor parameter θ_μ is trained by maximizing the policy gradient using the Adam optimizer

$$\theta_\mu = \underset{\theta_\mu}{\arg\max} \sum_{j=1}^{J} \nabla_{\mathbf{a}^{(j)}} Q\left(\mathbf{s}^{(j)}, \mathbf{a}^{(j)}|\theta_\mu\right) \times \nabla_{\theta_\mu} \mu\left(\mathbf{s}^{(j)}|\theta_\mu\right), \quad (27)$$

where $\nabla_{\mathbf{a}^{(j)}} Q\left(\mathbf{s}^{(j)}, \mathbf{a}^{(j)}|\theta_\mu\right)$ is the policy gradient of Q-function and $\nabla_{\theta_\mu} \mu\left(\mathbf{s}^{(j)}|\theta_\mu\right)$ is the policy gradient of the actor function. Target networks only receive a soft update after τ time slots to increase the stability of the algorithm, as mentioned above. In particular, the weights of target networks are updated by

$$\theta_{Q'_1} \leftarrow \lambda\theta_Q + (1-\lambda)\theta_{Q'_1}, \quad (28)$$

$$\theta_{Q'_2} \leftarrow \lambda\theta_Q + (1-\lambda)\theta_{Q'_2}, \quad (29)$$

and

$$\theta_{\mu'} \leftarrow \lambda\theta_\mu + (1-\lambda)\theta_{\mu'}. \quad (30)$$

where λ is the learning rate.

V. SIMULATION RESULTS AND DISCUSSIONS

In this section, the effectiveness of the proposed joint designs is illustrated through simulation results. A 3D Cartesian coordinate system whose origin is the floor center is used to specify the positions of LED luminaires, Bob and Eve. The dimension of the room is denoted as $d_L \times d_W \times d_H$, and the height of the receiver relative to the floor is set as d_R. Without being noted otherwise, the simulation parameters are given in Table 3. Figure 5 illustrates the room's general layout where there are 4 LED luminaires connected to a central processing unit (CPU) via wired connections. Moreover, simulation results are obtained by averaging the results corresponding to 100 randomly generated Bob's and Eve's channels.

Remark on the computational complexity: The complexity of the Q-learning-based joint design depends on the size of precoder $\mathbf{w}$, the number of quantized precoders $2T_s + 1$, numbers of possible modulation orders $\mathcal{M}$, and learning iterations k. In the case of the Deep Q-learning-based approach, the complexity of the joint design depends on the magnitude of the sampled experience S_{exp}, the length of the state sequence J, the number of filters in the critic network $n_1^f = 16$, $n_2^f = 32$, and the kernels $m_2 = 4$ in the critic network. Table 2 illustrates the unique utilized neural network setup. According to [46]–[48], the computational complexities of Q-learning-based and Deep Q-learning-based joint designs are thus $\mathcal{O}\left(kN(2T_s+1)\mathcal{M}\right)$ and $\mathcal{O}\left(S_{\text{exp}} J n_1^f n_2^f m_2\right)$, respectively.

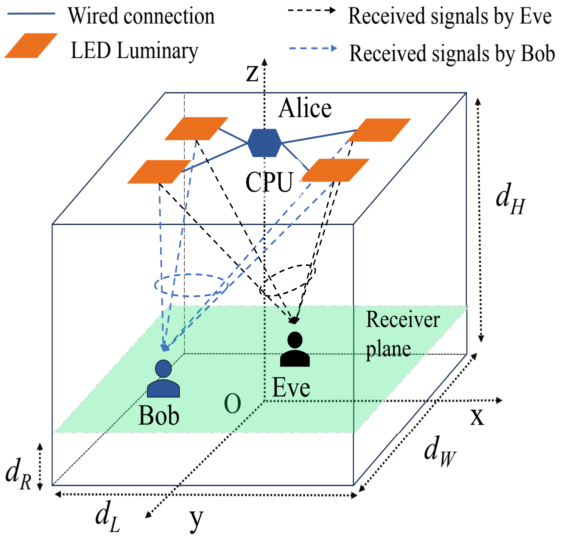
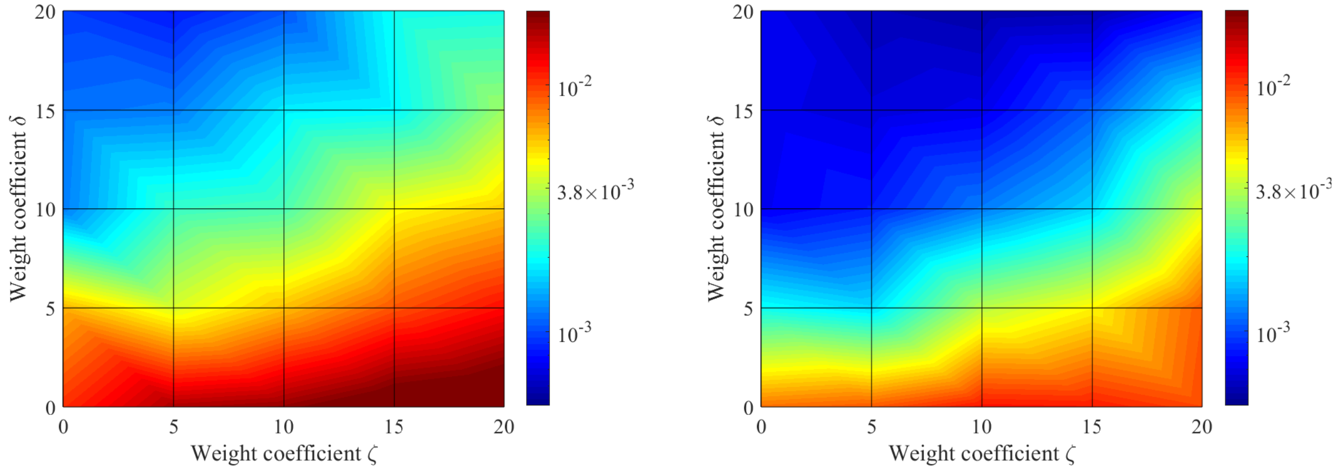


FIGURE 5: Geometrical configuration of the considered system.

TABLE 3: Simulation parameters.

Parameter	Value
Room Dimension	5 m × 5 m × 3 m
Positions of LED luminaires	$[-\sqrt{5}, -\sqrt{5}, 3]$ $[+\sqrt{5}, -\sqrt{5}, 3]$ $[+\sqrt{5}, +\sqrt{5}, 3]$ $[-\sqrt{5}, +\sqrt{5}, 3]$
Transmit power per luminaire, p_t	5W
LED beam angle, ϕ	120°
LED conversion factor, η	0.44 W/A
PD active area, A_r	1 cm^2
PD responsivity, γ	0.54 A/W
PD field of view (FoV), Ψ	60°
Optical filter gain, $T_s(\psi)$	1
Refractive index of concentrator, κ	1.5
Modulation index, α	0.1
Average noise power, σ_{R}^2	-98.82 dBm
Learning rate, λ	0.5
Discount parameter, β	0.5
Batches size, J	128
Update frequency, τ	4

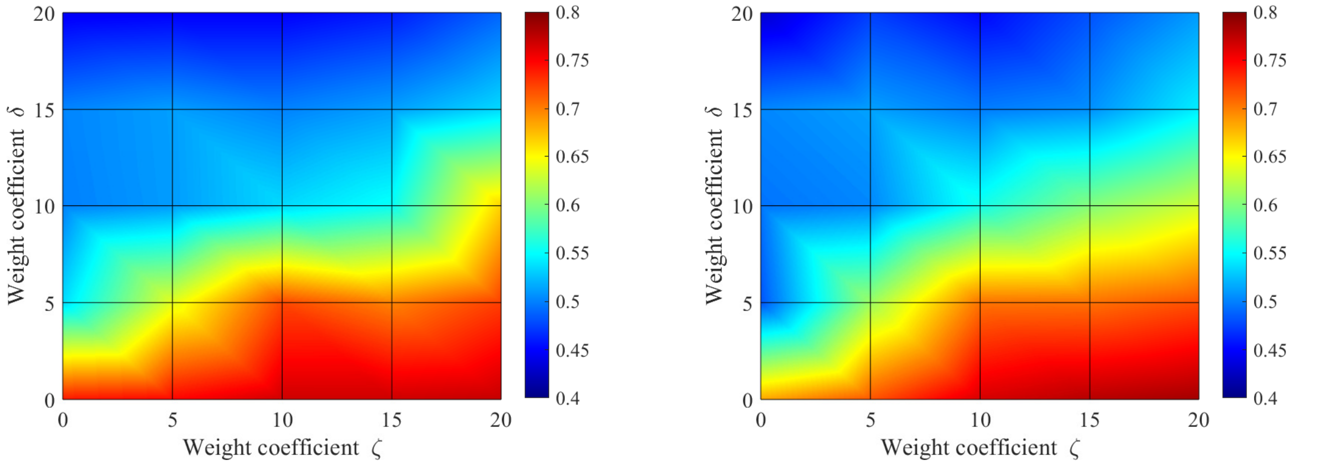
Given the CSI of Bob's and Eve's channels (i.e., the location of Bob and Eve), the optimal precoder and modulation order depend solely on the formulation of the reward func-



(a) Q-learning

(b) Deep Q-learning

FIGURE 6: Impact of δ and ζ on the BER of Bob's channel.



(a) Q-learning

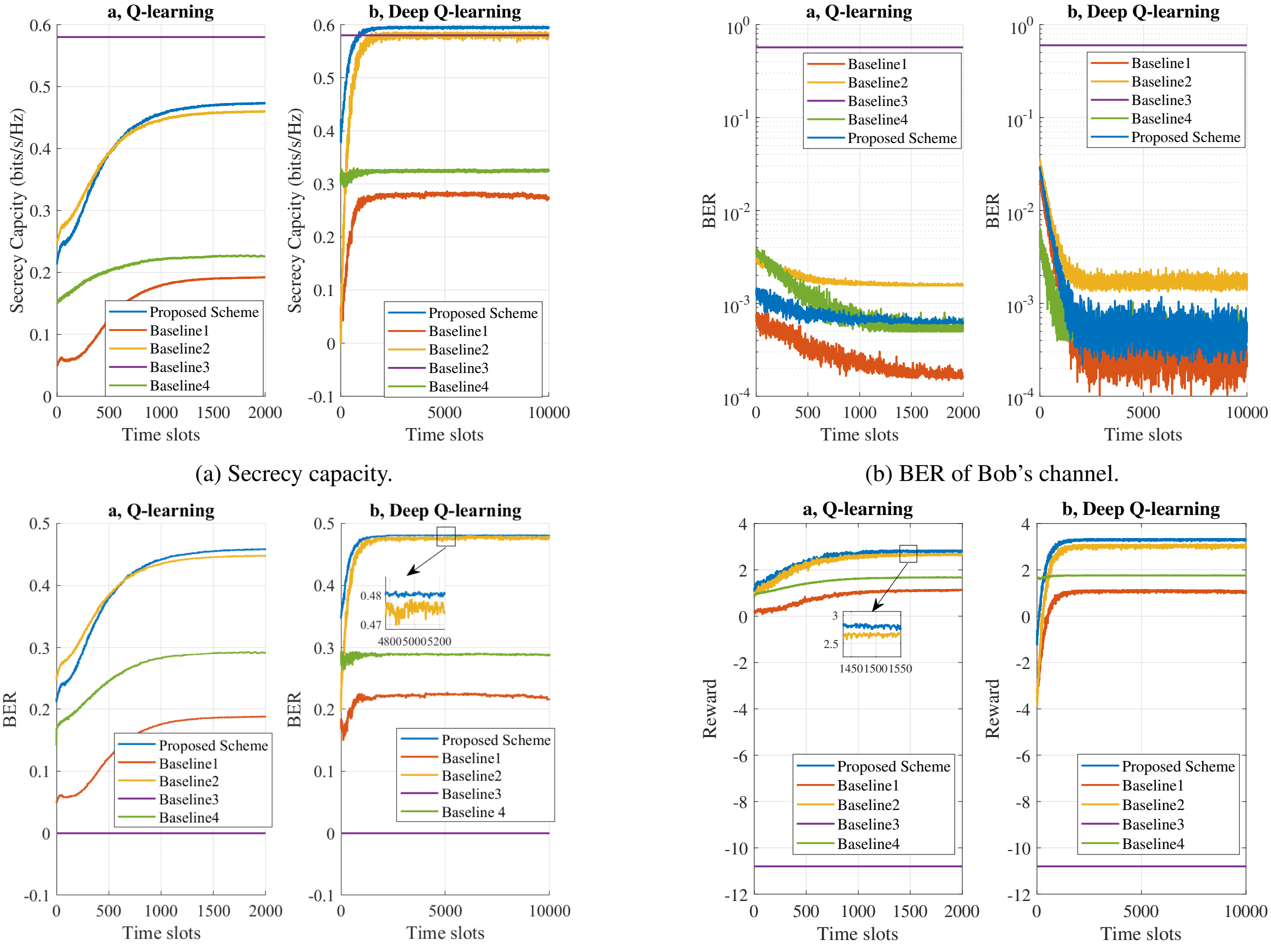
(b) Deep Q-learning

FIGURE 7: Impact of δ and ζ on the BER of Eve's channel.

tion, specifically the choice of the weight coefficients δ and ζ in the reward function. Note that the BERs of Bob's and Eve's channels are functions of the precoder and modulation order. Since one of the objectives of the proposed joint design is to ensure the reliability of Bob's channel and the unreliability of Eve's channel (which are determined by their BER values), it is necessary to investigate the impact of adjusting δ and ζ on the BERs of Bob's and Eve' channel. Such influence is illustrated in Figs. 6a, 6b for Bob's channel and Figs. 7a, 7b for Eve's channel, respectively. Here, we run the learning algorithms for different values of δ and ζ. The resulting BERs of Bob's and Eve's channels are then calculated using the obtained optimal modulation order and precoder. We notice that the BERs of Bob's and Eve's channels are both proportional to ζ and inversely proportional to δ. Hence, one can lower the BER of Bob's channel (thus guaranteeing the reliability of its channel) by choosing a large δ and a small ζ. An intuitive explanation for this behavior is that when δ decreases and/or ζ increases, to keep maximizing the reward defined in (23), the BER of Bob's channel $p_{e,\mathrm{B}}$ is *forced* to decrease as well. Note that the reduction of the BER of Bob's channel is often the result of lower modulation order. This explains the same behavior of the BER of Eve's channel (i.e., the proportionality and inverse proportionality with respect to ζ and δ, respectively). In this work, the pre-FEC BER threshold for asserting the channel reliability is chosen at the common value of 3.8×10^{-3} [49], [50]. Accordingly, to comfortably satisfy this threshold (i.e., the BER of Bob's (Eve's) channel is lower (higher) than 3.8×10^{-3}), $(\delta,\ \zeta) = (20, 5)$ is chosen for subsequent simulations.

To highlight the effectiveness of the proposed design, four different baseline schemes described in the following are considered for comparison.

- **Baseline 1**: The modulation is fixed at the lowest order



(a) Secrecy capacity.

(b) BER of Bob's channel.

(c) BER of Eve's channel.

(d) Reward value.

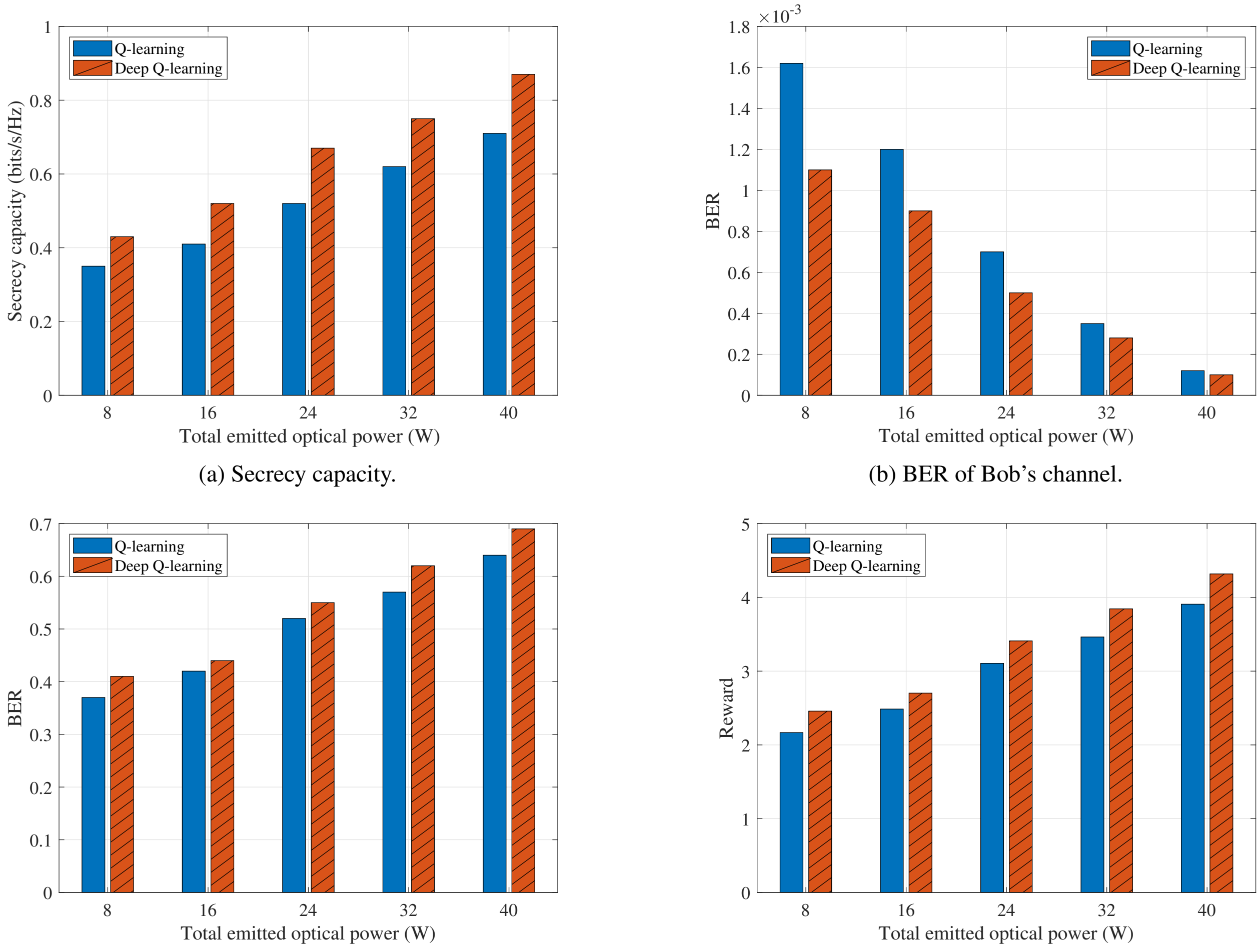
FIGURE 8: Comparisons of proposed joint design with baseline schemes.

2-PAM. This baseline design prioritizes a low BER (i.e., high channel reliability) at the expense of reduced secrecy capacity.

- **Baseline 2**: The modulation is fixed at the higher order 64-PAM. In contrast to **Baseline 1**, this baseline aims for a high secrecy capacity at the expense of high BER (i.e., low channel reliability).
- **Baseline 3**: The precoder $\mathbf{w}$ lies on the null-space of $\mathbf{h}_{\mathrm{E}}^T$, i.e., $\mathbf{h}_{\mathrm{E}}^T\mathbf{w} = 0$. The design principle of this baseline is to ensure that Eve cannot receive any signal. The secrecy capacity is, therefore, the capacity of Bob's channel.
- **Baseline 4**: We compare our proposed joint design with that proposed in [27] where a modulation-independent lower bound secrecy capacity formula (given in [28]) and an approximate 4PAM BER expression was used.

Performances in terms of the secrecy capacity, BERs of Bob's and Eve's channels, and reward are respectively shown in Figs. 8a, 8b, 8c, and 8d for the proposed joint design and baselines. Here, comparisons between the Q-learning-based and Deep Q-learning-based approaches are also illustrated. For the secrecy capacity, the proposed DRL-based design achieves the highest performance, which is double that of the **Baseline 1 and Baseline 4** and slightly better than those of the **Baseline 2** and **3**. However, in the case of Q-learning-based design, the performance of **Baseline 3** is considerably better than that of the proposed joint design at the expense of very high BER of Bob's channel (i.e., $\sim$ 8 $\times$ 10^{-1} which is significantly above the pre-FEC BER threshold). It is verified that the proposed joint design, **Baseline 1, 2** and **4** all satisfy the pre-FEC BER threshold of Bob's channel (i.e., BER of Bob's channel is less than 3.8×10^{-3}) while rendering the BER of Eve's channel significantly higher than the threshold. We note here that in the case of **Baseline 3**, as Eve cannot receive any signal (due to $\mathbf{h}_{\mathrm{E}}^T\mathbf{w} = 0$), the BER of Eve's channel is defined as 0 by convention[4]. **Baseline 4** performs slightly better than Baseline 1, as it achieves a higher secrecy capacity while maintaining satisfactory BER of Bob's channel. It is, however, inferior to our proposed design, which achieves almost twice the reward and secrecy capacity. Overall, taking into account the secrecy capacity

[4] As a consequence, the reward function for the **Baseline 3** is defined by $u = C_s - \delta p_{e,\mathrm{B}}$.



(a) Secrecy capacity.

(b) BER of Bob's channel.

(c) BER of Eve's channel.

(d) Reward value.

FIGURE 9: Comparison between the performance of Q-learning-based and Deep Q-learning-based designs.

and the BERs of Bob's and Eve's channels, the reward value displayed in Fig. 8d shows the superiority of the proposed design over the four baseline schemes.

We compare in Figs. 9a, 9b, 9c, and 9d the performance of Q-learning-based and Deep Q-learning-based designs with respective to the total emitted optical power (i.e., the sum optical power of all luminaires). Firstly, the superiority of the Deep Q-learning-based approach due to its ability to tackle continuous action space (thus being able to select the optimal action) is clearly illustrated. For example, at the total transmit power of 16 W (i.e., 4 W for each luminaire), the deep Q-learning-based design achieves 25.4% better secrecy capacity, 25% lower BER of Bob's channel, 4.7% higher BER of Eve's channel, and thus resulting in 8.6% higher reward value than the Q-learning-based design does. While it is relatively obvious that an increase in the total emitted optical power results in an increased secrecy capacity and a decreased BER of Bob's channel (as illustrated in Figs. 9a and 9b), it is interesting that the BER of Eve's channel increases with the optical power. This is desirable as Eve's channel's unreliability is further guaranteed. Through observing simulation results (i.e., optimal modulation order and precoder), it is revealed that as the optical power increases, the learning algorithms tend to select higher modulation orders to maximize the reward value since high modulation order leads to higher secrecy capacity. Note that although an increased modulation order would worsen the BER performance, in the case of Bob's channel, the combined effect of increased power and proper precoder (the precoder is designed for the sake of keeping the BER of Bob's channel as low as possible) dominates that negative impact, thus resulting in a decreased BER performance. In the case of Eve's channel, however, the effect of increased power alone is not enough to keep the BER decreasing as the modulation order increases. Compared with Bob's channel, we, thus, observe this contrasting behavior of the BER of Eve's channel.

VI. CONCLUSION

In this paper, a joint design that integrates precoding and adaptive M-PAM modulation was proposed for the wiretap MISO VLC channels. The proposed design aimed to maximize the secrecy capacity while simultaneously ensuring the reliability of Bob's channel and the unreliability of Eve's channel. Under the constraint on the signal amplitude, reinforcement learning approaches based on Q-learning and Deep Q-learning were proposed to jointly optimize the precoder and the modulation order. To speed up the learning process, we also presented a high-accuracy yet low-complexity approximation to the secrecy capacity. Simulation results revealed that by properly choosing the weight factors for the reward function, the BERs of Bob's and Eve's channels could be made respectively lower and higher the pre-FEC BER threshold, thus satisfying the reliability and unreliability requirements. It was also shown that the proposed joint design outperformed the considered baseline schemes. Note that the position of Eve is often not known in practice as she tries to hide her presence from Alice. Therefore, our future work aims at joint adaptive modulation and precoding designs for such a scenario. Moreover, the receiver's orientation and channel blockage are two critical impairment factors in practical VLC systems. Thus, it would be interesting to investigate the performance of the proposed joint design considering these effects.

REFERENCES

[1] D. M. T. Hoang, T. V. Pham, A. T. Pham, and C. T. Nguyen, "Q-learning-based joint design of adaptive modulation and precoding for physical layer security in visible light communications," in 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), pp. 1–5, 2023.

[2] S. Rajagopal, R. D. Roberts, and S.-K. Lim, "Ieee 802.15. 7 visible light communication: modulation schemes and dimming support," IEEE Communications Magazine, vol. 50, no. 3, pp. 72–82, 2012.

[3] "IEEE standard for local and metropolitan area networks–part 15.7: Short-range wireless optical communication using visible light," IEEE Std 802.15.7-2011, pp. 1–309, 2011.

[4] L. Yin and H. Haas, "Physical-layer security in multiuser visible light communication networks," IEEE Journal on Selected Areas in Communications, vol. 36, no. 1, pp. 162–174, 2018.

[5] S. Arnon, Visible light communication. Cambridge University Press, 2015.

[6] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," IEEE communications surveys & tutorials, vol. 17, no. 4, pp. 2047–2077, 2015.

[7] F. Wang, F. Yang, J. Song, and Z. Han, "Access frameworks and application scenarios for hybrid vlc and rf systems: state of the art, challenges, and trends," IEEE Communications Magazine, vol. 60, no. 3, pp. 55–61, 2022.

[8] G. Blinowski, "Security of visible light communication systems-a survey," Physical Communication, vol. 34, pp. 246–260, 2019.

[9] A. D. Wyner, "The wire-tap channel," Bell system technical journal, vol. 54, no. 8, pp. 1355–1387, 1975.

[10] I. Csiszár and J. Korner, "Broadcast channels with confidential messages," IEEE Transactions on Information Theory, vol. 24, no. 3, pp. 339–348, 1978.

[11] M. A. Arfaoui, M. D. Soltani, I. Tavakkolnia, A. Ghrayeb, M. Safari, C. M. Assi, and H. Haas, "Physical layer security for visible light communication systems: A survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1887–1908, 2020.

[12] O. Ozel, E. Ekrem, and S. Ulukus, "Gaussian wiretap channel with amplitude and variance constraints," IEEE Transactions on Information Theory, vol. 61, no. 10, pp. 5553–5563, 2015.

[13] J.-Y. Wang, C. Liu, J.-B. Wang, Y. Wu, M. Lin, and J. Cheng, "Physical-layer security for indoor visible light communications: Secrecy capacity analysis," IEEE Transactions on Communications, vol. 66, no. 12, pp. 6423–6436, 2018.

[14] J.-Y. Wang, X.-T. Fu, R.-R. Lu, J.-B. Wang, M. Lin, and J. Cheng, "Tight capacity bounds for indoor visible light communications with signal-dependent noise," IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1700–1713, 2021.

[15] J.-Y. Wang, P.-F. Yu, X.-T. Fu, J.-B. Wang, M. Lin, J. Cheng, and M.-S. Alouini, "Secrecy-capacity bounds for visible light communications with signal-dependent noise," IEEE Transactions on Wireless Communications, pp. 1–1, 2023.

[16] A. Mostafa and L. Lampe, "Optimal and robust beamforming for secure transmission in MISO visible-light communication links," IEEE Transactions on Signal Processing, vol. 64, no. 24, pp. 6501–6516, 2016.

[17] S. Ma, Z.-L. Dong, H. Li, Z. Lu, and S. Li, "Optimal and robust secure beamformer for indoor MISO visible light communication," Journal of Lightwave Technology, vol. 34, no. 21, pp. 4988–4998, 2016.

[18] T. V. Pham and A. T. Pham, "Secrecy sum-rate of multi-user MISO visible light communication systems with confidential messages," Optik, vol. 151, pp. 65–76, 2017.

[19] M. A. Arfaoui, A. Ghrayeb, and C. M. Assi, "Secrecy performance of multi-user MISO VLC broadcast channels with confidential messages," IEEE Transactions on Wireless Communications, vol. 17, no. 11, pp. 7789–7800, 2018.

[20] S. Cho, G. Chen, and J. P. Coon, "Zero-forcing beamforming for active and passive eavesdropper mitigation in visible light communication systems," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1495–1505, 2021.

[21] S. T. Duong, T. V. Pham, C. T. Nguyen, and A. T. Pham, "Energy-efficient precoding designs for multi-user visible light communication systems with confidential messages," IEEE Transactions on Green Communications and Networking, vol. 5, no. 4, pp. 1974–1987, 2021.

[22] H. Shen, Y. Deng, W. Xu, and C. Zhao, "Secrecy-oriented transmitter optimization for visible light communication systems," IEEE Photonics Journal, vol. 8, no. 5, pp. 1–14, 2016.

[23] S. Cho, G. Chen, and J. P. Coon, "Enhancement of physical layer security with simultaneous beamforming and jamming for visible light communication systems," IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2633–2648, 2019.

[24] T. V. Pham, T. Hayashi, and A. T. Pham, "Artificial-noise-aided precoding design for multi-user visible light communication channels," IEEE Access, vol. 7, pp. 3767–3777, 2019.

[25] T. V. Pham and A. T. Pham, "Energy efficient artificial noise-aided precoding designs for secured visible light communication systems," IEEE Transactions on Wireless Communications, vol. 20, no. 1, pp. 653–666, 2020.

[26] A. Alvarado, E. Agrell, D. Lavery, R. Maher, and P. Bayvel, "Replacing the soft-decision FEC limit paradigm in the design of optical communication systems," Journal of Lightwave Technology, vol. 34, no. 2, pp. 707–721, 2016.

[27] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," IEEE transactions on communications, vol. 67, no. 10, pp. 6994–7005, 2019.

[28] A. Mostafa and L. Lampe, "Physical-layer security for MISO visible light communication channels," IEEE Journal on Selected Areas in Communications, vol. 33, no. 9, pp. 1806–1818, 2015.

[29] S. Rezaei Aghdam, A. Nooraiepour, and T. M. Duman, "An overview of physical layer security with finite-alphabet signaling," IEEE Commun. Surv. Tutor., vol. 21, no. 2, pp. 1829–1850, 2019.

[30] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," IEEE Trans. Consum. Electron., vol. 50, no. 1, pp. 100–107, 2004.

[31] M. Abd Elkarim, M. Elsherbini, H. M. AbdelKader, and M. H. Aly, "Exploring the effect of led nonlinearity on the performance of layered aco-ofdm," Applied Optics, vol. 59, no. 24, pp. 7343–7351, 2020.

[32] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," IEEE Signal Processing Magazine, vol. 31, no. 4, pp. 142–148, 2014.

[33] A. Papoulis, Probability, random variables, and stochastic processes. McGraw-Hill, 1991.

[34] K. Cho and D. Yoon, "On the general ber expression of one- and two-dimensional amplitude modulations," IEEE Transactions on Communications, vol. 50, no. 7, pp. 1074–1080, 2002.

[35] E. Illi, E. Baccour, M. Qaraqe, and M. Hamdi, "Deep reinforcement learning for enhancing the secrecy of a MU-MISO UOWC network," in GLOBECOM 2023 - 2023 IEEE Global Communications Conference, pp. 6807–6812, 2023.

[36] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination," IEEE Transactions on Communications, vol. 68, no. 10, pp. 6070–6085, 2020.

[37] X. Zhou, X. Zhang, C. Chen, Y. Niu, Z. Han, H. Wang, C. Sun, B. Ai, and N. Wang, "Deep reinforcement learning coordinated receiver beamforming for millimeter-wave train-ground communications," IEEE Transactions on Vehicular Technology, vol. 71, no. 5, pp. 5156–5171, 2022.

[38] C. J. C. H. Watkins, Learning from Delayed Rewards. PhD thesis, King's College, Oxford, 1989.

[39] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, pp. 279–292, May 1992.

[40] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, pp. 529–533, Feb. 2015.

[41] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in International conference on machine learning, pp. 1587–1596, PMLR, 2018.

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

[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), (San Diega, CA, USA), 2015.

[44] G. Leuenberger and M. A. Wiering, "Actor-critic reinforcement learning with neural networks in continuous games," in ICAART 2018-Proceedings of the 10th International Conference on Agents and Artificial Intelligence, SciTePress, 2018.

[45] M. J. Peixoto and A. Azim, "Using time-correlated noise to encourage exploration and improve autonomous agents performance in reinforcement learning," Procedia Computer Science, vol. 191, pp. 85–92, 2021.

[46] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?," Advances in neural information processing systems, vol. 31, 2018.

[47] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5353–5360, 2015.

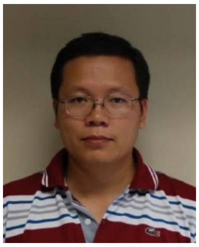
[48] L. Xiao, X. Lu, T. Xu, X. Wan, W. Ji, and Y. Zhang, "Reinforcement learning-based mobile offloading for edge computing against jamming and interference," IEEE Transactions on Communications, vol. 68, no. 10, pp. 6114–6126, 2020.

[49] C.-W. Chow et al., "Actively controllable beam steering optical wireless communication (OWC) using integrated optical phased array (OPA)," Journal of Lightwave Technology, vol. 41, no. 4, pp. 1122–1128, 2023.

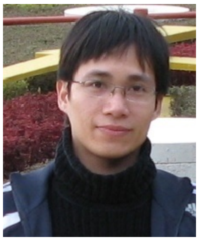
[50] J. Zhang et al., "PAPR reduction and nonlinearity mitigation of optical digital subcarrier multiplexing systems with a silicon photonics transmitter," Journal of Lightwave Technology, vol. 41, no. 22, pp. 6957–6969, 2023.

DUC M. T. HOANG received a B.E. degree in Electronics and Communications from the Hanoi University of Science and Technology (HUST) in 2023. He is currently a Research Assistant at the Communication Theory and Applications Research Group, School of Electrical and Electronic Engineering, HUST. His research interests include physical layer security and visible light communications.

THANH V. PHAM (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science and engineering from the University of Aizu, Japan, in 2014, 2016, and 2019, respectively. From April 2018 to March 2020, he was a Japan Society for the Promotion of Science (JSPS) Fellow. He was a Special Postdoctoral Fellow at the School of Computer Science and Engineering, University of Aizu, from April 2020 to March 2021, and a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, McMaster University, Canada, from April 2021 to February 2022. He has been an Assistant Professor at the Department of Mathematical and Systems Engineering, Shizuoka University, Japan, since March 2022. His research interests include optical wireless communications and communication theory. He is also a member of IEICE and IPSJ.



ANH T. PHAM (Senior Member, IEEE) received the B.E. and M.E. degrees in electronics engineering from the Hanoi University of Technology, Vietnam, in 1997 and 2000, respectively, and the Ph.D. degree in information and mathematical sciences from Saitama University, Japan, in 2005. From 1998 to 2002, he was with NTT Corporation, Vietnam. Since 2005, he has been a Faculty Member at The University of Aizu, where he is currently a Professor and the Head of the Computer Communications Laboratory, Division of Computer Engineering. His research interests are in the broad areas of communication theory and networking with a particular emphasis on modeling, design, and performance evaluation of wired/wireless communication systems and networks. He has authored/co-authored over 160 peer-reviewed papers on these topics. Dr. Pham is a senior member of IEEE. He is also a member of IEICE and OSA.



CHUYEN T. NGUYEN received his B.E. degree in Electronics and Telecommunications, Hanoi University of Science and Technology (HUST), Vietnam in 2006, M.S. degree in Communications Engineering from National Tsing-Hua University, Taiwan in 2008, and a Ph.D. degree in Informatics from Kyoto University, Japan in 2013. From September to November 2014, he was a visiting researcher at The University of Aizu, Japan. He received the Fellow award from the Hitachi Global Foundation in August 2016, the First Best Paper Award in 2019 IEEE ICT, the Best Paper Award in 2018 KICS/IEEE ICTC, and the 2019 IEEE ICC. He is currently an Associate Professor in the School of Electrical and Electronic Engineering, HUST, Vietnam. His current research interests are in areas of communication theory and applications, with particular emphasis on protocol design for industrial Internet of Things applications and wireless/optical networks.