基于深度强化学习的可见光定位通信一体化功率分配研究

马帅，李兵，盛海鸿，谷荣妍，周辉，王洪梅，王悦，李世银
（中国矿业大学信息与控制工程学院，江苏 徐州 221000）

摘 要：为了实现定位通信一体化功率分配，提出了一种基于深度强化学习的可见光定位通信（VLPC）一体化系统的功率分配方案。首先，提出了定位通信一体化帧结构设计；其次，利用定位信息实现了信道状态信息的估计，并推导了定位误差的克拉美罗下界（CRLB）；再次，阐明了定位精度和通信速率的内在耦合关系；最后，提出了基于深度确定性策略梯度的 VLPC 动态功率分配方案。仿真结果表明，所提方案可同时实现高精度定位和高速通信。

关键词：可见光定位通信一体化；克拉美罗下界；功率分配；强化学习
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022163

Research on power allocation of integrated VLPC based on deep reinforcement learning

MA Shuai, LI Bing, SHENG Haihong, GU Rongyan, ZHOU Hui,WANG Hongmei, WANG Yue, LI Shiyin
School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221000, China

Abstract: A power allocation scheme for integrated visible light position and communication (VLPC) system based on deep reinforcement learning was proposed to achieve power allocation for communication positioning integration. First, the frame structure design of integrated VLPC was proposed. Then the channel state information could be estimated by using the positioning information, and the CRLB of the positioning error was derived. Furthermore, the internal coupling relationship between positioning accuracy and communication rate was clarified. On this basis, a dynamic power allocation scheme based on deep deterministic policy gradient was proposed. Simulation results show that the proposed scheme can simultaneously achieve high-precision positioning and high-speed communication.

Keywords: integrated VLPC, Cramér-Rao lower bound, power allocation, reinforcement learning

0 引言

随着移动互联网业务不断深入，无线设备数量的增长和高速通信需求的增多给传统的通信网络带来了很大压力[1]。据 Cisco 预计，2023 年全球互联网用户规模将达到 53 亿，移动设备总数将达到 131 亿，且值得注意的是，超过 50%的语音流量和 70%的无线数据流量发生在室内环境，这使射频（RF, radio frequency）通信频谱资源短缺问题日益突出。可见光通信（VLC, visible light communication）和可见光定位（VLP, visible light position）由于具有远高于 RF 频段的免授权带宽（430～790 THz）、绿色节能和电磁免疫等优点，近年来受到学术界和工业界的研究关注[2-3]。

收稿日期：2022–06–10；**修回日期**：2022–08–08
通信作者：李世银，lishiyin@cumt.edu.cn
基金项目：中央高校基本科研业务费专项资金资助项目（No.2022QN1052）；江苏省自然科学基金资助项目（No.BK20221115）；中国矿业大学未来科学家计划基金资助项目（No.2022WLJCRCZL108）
Foundation Items: The Fundamental Research Funds for the Central Universities (No.2022QN1052), The Natural Science Foundation of Jiangsu Province (No.BK20221115), Funded by the Graduate Innovation Program of China University of Mining and Technology (No.2022WLJCRCZL108)

VLC 通常利用简单的强度调制和直接检测方式进行信息传输，可同时实现照明和通信功能，被视为 B5G 和 6G 的关键技术之一[4]。目前已有大量文献研究如何提升 VLC 系统的性能。文献[5]研究了多用户 VLC 系统的容量区域，并提出了一种基于交替方向乘子法的干扰管理方案。文献[6]研究了无小区 VLC 系统的资源分配方案，可在满足光功率约束和用户速率要求的条件下提高通信速率。文献[7]利用五色硅基 LED，实现了速率为 14.6 Gbit/s 的水下 VLC 系统。此外，作为 VLC 的一个重要应用，VLP 能实现室内高精度定位，可应用于室内导航和物流管理等领域中。通过测量不同的可见光信号特性，现有的 VLP 方案主要使用基于接收信号强度指示（RSSI, received signal strength indication）、到达角度、到达时间和到达时间差等技术。在上述方案中，RSS 定位技术由于其复杂度较低和部署简单等优势被广泛采用。文献[8]使用 RSS 和接收端多个光电二极管（PD, photodiode）之间的相对位置来确定目标的位置，在对称的圆形区域实现了定位精度小于 1.5 cm 的定位误差。文献[9]研究了由单个 LED 和多个倾斜接收器组成的室内高精度 VLP 系统，基于倾斜角的先验信息，使用 RSS 定位方法进行二维和三维的定位，定位精度小于 6 cm。文献[10]提出了使用机器学习方法提升基于 RSS 的 VLP 系统的定位精度的方案，并通过线性插值减少了模型训练所需的样本数。

现有研究大多仅关注 VLC 或 VLP 系统，而在室内场景下关于可见光定位通信（VLPC, visible light position and communication）一体化系统资源分配的研究具有很强的现实意义，近年来受到了研究者的重视。文献[11]研究了一个面向物联网的 VLPC 一体化系统的资源分配问题，在保证定位精度约束的同时，通过联合优化用户接入、带宽分配和功率分配以最大化传输速率。文献[12]基于正交频分复用（OFDM, orthogonal frequency division multiplexing）技术，提出了一种适用于室内可见光通信和定位的稳健传输方案，通过将 LED 的发射信号调制到不同的子载波以克服用户间干扰，仿真结果表明该方案能同时满足通信和定位要求。文献[13]提出了基于滤波器组多载波调制技术的 VLPC 一体化系统传输方案，相较于 OFDM，该方案可有效提高带宽利用率。文献[14]考虑到 VLPC 一体化系统中定位精度和用户最低通信速率要求，提出了一种基于无模型强化学习的资源分配方案来最大化多用户和速率。

在实际的 VLPC 系统中，由于用户的移动性、不准确的信道状态信息（CSI, channel state information）和对服务质量的要求，导致很难获得动态系统的完整信息，使传统优化方法很难解决该类具有时变特征的优化问题。近年来，深度强化学习（DRL, deep reinforcement learning）被广泛应用于多种复杂无线通信环境下的动态资源分配问题中，其主要思想是在与环境的长期交互过程中，利用深度学习感知环境，利用强化学习改善策略[15-16]。为了解决上述问题，本文旨在研究 VLPC 一体化系统的基本原理，并提出一种基于深度确定性策略梯度（DDPG, deep deterministic policy gradient）的动态功率分配方案，主要的研究工作如下。

1) 建立了移动用户场景下 VLPC 一体化系统模型，通过帧结构的设计，使发射端不需要利用导频序列进行信道状态信息估计，而是根据定位信息获得该结果，这可以显著降低系统开销；推导了定位误差的克拉美罗下界（CRLB, Cramér-Rao lower bound）和可达通信速率的表达式，揭示定位和通信的内在关系。

2) 研究了满足 CRLB 门限、实际光功率约束和总功率约束条件下的动态功率分配问题，以最大化移动用户的平均速率。由于该问题难以用传统优化方法解决，首先将该问题重构为马尔可夫决策过程，然后提出一种基于 DDPG 的动态功率分配算法，以充分发掘历史数据中有价值的信息。

3) 仿真结果表明，本文所提算法能取得良好的通信性能，并能缓解定位误差带来的影响。与深度 Q 网络（DQN, deep Q network）和等功率分配方案对比，验证了本文算法的有效性。

1 VLPC 一体化系统模型

考虑一个室内下行链路 VLPC 一体化系统，如图 1 所示，包括一个配备 N 个 LED 的发射基站和一个配备单个 PD 的移动用户。定义 LED 的索引集为 $\mathcal{N} \triangleq \{1,2,\cdots,N\}$，第 i 个 LED 的位置为 $\boldsymbol{v}_i = [x_i, y_i, z_i]^{\mathrm{T}}, \forall i \in \mathcal{N}$，在时隙 t 处用户的位置为 $\boldsymbol{u}(t) = [x_{\mathrm{u}}(t), y_{\mathrm{u}}(t), z_{\mathrm{u}}(t)]^{\mathrm{T}}$，且信道状态在每个时隙内保持不变。

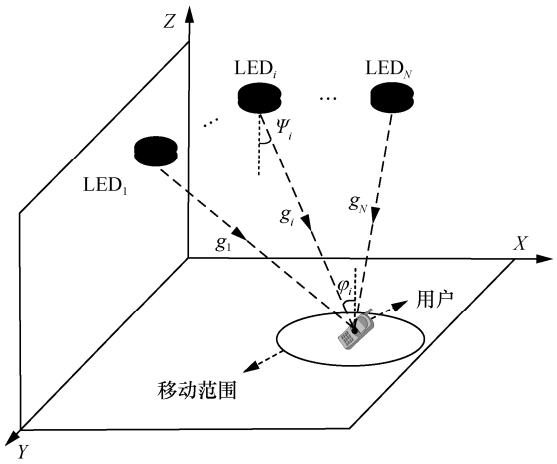


图 1　VLPC 一体化系统模型

如图 2 所示，在每个时隙上，VLPC 一体化系统发送的信息帧被划分为定位子帧、反馈子帧和通信子帧。具体地，在定位子帧中，基站先向用户发送定位信号，接收端则通过接收到信号的 RSS 值来估计用户位置。在反馈子帧中，用户的估计位置被反馈给基站，基站再根据估计位置计算每个 LED 与用户之间的 CSI 估计值。在通信子帧中，基站根据 CSI 估计值与用户进行定向通信。

定位子帧	反馈子帧	通信子帧

图 2　VLPC 一体化系统帧结构

考虑到 LED 与用户之间的信道增益取决于直射链路[17]，定义时隙 t 处第 i 个 LED 和移动用户之间的信道增益为

$$g_i(t)=\frac{(m+1)A_r}{2\pi d_i^2(t)}\cos^m\left(\psi_i\right)\cos\left(\varphi_i\right)F_s\Gamma\left(\varphi_i\right) \quad (1)$$

其中，$m=-\dfrac{\ln 2}{\ln\left(\cos\left(\psi_{\frac{1}{2}}\right)\right)}$ 表示朗伯指数，$\psi_{\frac{1}{2}}$ 表示 LED 的半功率角，A_r 表示 PD 的感光面积，$d_i\left(t\right)$ 表示第 i 个 LED 到用户的距离，ψ_i 表示第 i 个 LED 的辐射角，φ_i 表示 PD 的入射角，F_s 表示接收端滤光片的增益，$\Gamma\left(\varphi_i\right)$ 表示聚光器的增益。当 $0\leqslant\varphi_k\leqslant\varphi_{\mathrm{FOV}}$ 时，$\Gamma\left(\varphi_i\right)=\dfrac{n_r^2}{\sin^2\left(\varphi_{\mathrm{FOV}}\right)}$；否则 $\Gamma\left(\varphi_k\right)=0$，其中，$n_r$ 表示折射率，φ_{FOV} 表示视场角。

1.1　定位子帧

在定位子帧中，基站先发送定位信号给用户，接收端根据获得的 RSS 值计算用户的估计位置。具体地，在时隙 t 处，定义 $s_{\mathrm{p},i}(t)$ 表示第 i 个 LED 发送的定位光信号，满足峰值约束 $-A\leqslant s_{\mathrm{p},i}\left(t\right)\leqslant A$，均值约束 $\mathbb{E}\left\{s_{\mathrm{p},i}(t)\right\}=0$，均方约束 $\mathbb{E}\left\{s_{\mathrm{p},i}^2(t)\right\}=\varepsilon$，其中，$A>0$ 和 ε 分别表示光信号的峰值和方差。第 i 个 LED 发送的定位信号 $x_{\mathrm{p},i}(t)$ 为

$$x_{\mathrm{p},i}(t)=\sqrt{P_{\mathrm{p},i}(t)}s_{\mathrm{p},i}(t)+b,\forall i\in\mathcal{N} \quad (2)$$

其中，$P_{\mathrm{p},i}(t)$ 表示分配给第 i 个 LED 的定位功率，b 表示 LED 的直流偏置。为了保证发送信号的非负性，定位功率应满足 $\sqrt{P_{\mathrm{p},i}t}A\leqslant b$。

假设 LED 和 PD 指向均垂直于水平面，有 $\cos\left(\phi_i\right)=\cos\left(\varphi_i\right)=\dfrac{\left(z_{\mathrm{u}}(t)-z_i\right)}{\left\|\boldsymbol{u}(t)-\boldsymbol{v}_i\right\|_2}$，则式(1)可重写为

$$g_i(t)=\frac{\eta\left(z_{\mathrm{u}}(t)-z_i\right)^{m+1}}{\left\|\boldsymbol{u}(t)-\boldsymbol{v}_i\right\|_2^{m+3}},\forall i\in\mathcal{N} \quad (3)$$

其中，$\eta=\dfrac{(m+1)A_rF_s\Gamma\left(\varphi_i\right)}{2\pi}$。

移动用户在时隙 t 处接收到来自第 i 个 LED 的定位信号 $y_{\mathrm{p},i}(t)$ 可表示为

$$y_{\mathrm{p},i}(t)=g_i(t)x_{\mathrm{p},i}(t)+n_{\mathrm{p},i},\forall i\in\mathcal{N} \quad (4)$$

其中，$n_{\mathrm{p},i}$ 表示服从均值为零和方差为 σ_{p}^2 的高斯白噪声。

根据式(4)，用户接收到来自第 i 个 LED 的信号的电功率为

$$P_{\mathrm{rec},i}(t)=\mathbb{E}\left\{y_{\mathrm{p},i}^2(t)\right\}=\left(P_{\mathrm{p},i}(t)+b^2\right)g_i^2(t)+\sigma_{\mathrm{p}}^2 \quad (5)$$

联立式(3)和式(5)，可得到时隙 t 处关于用户位置的等式，表示为

$$\begin{cases}\dfrac{P_{\mathrm{rec},1}(t)-\sigma_{\mathrm{p}}^2}{\eta\left(P_{\mathrm{p},1}(t)+b^2\right)}=\dfrac{\left(z_1-z_{\mathrm{u}}(t)\right)^{m+1}}{\left\|\boldsymbol{v}_1-\boldsymbol{u}(t)\right\|_2^{m+3}}\\ \vdots\\ \dfrac{P_{\mathrm{rec},N}(t)-\sigma_{\mathrm{p}}^2}{\mu\left(P_{\mathrm{p},N}(t)+b^2\right)}=\dfrac{\left(z_N(t)-z_{\mathrm{u}}(t)\right)^{m+1}}{\left\|\boldsymbol{v}_N-\boldsymbol{u}(t)\right\|_2^{m+3}}\end{cases} \quad (6)$$

通过最小二乘法求解式(6)，可以得到时隙 t 处的用户估计位置 $\hat{\boldsymbol{u}}(t)=\left[\hat{x}_{\mathrm{u}}(t),\hat{y}_{\mathrm{u}}(t),\hat{z}_{\mathrm{u}}(t)\right]^{\mathrm{T}}$。

从实际角度看，由于噪声和非视距传输等因素的影响，定位误差难以避免，令 $\boldsymbol{e}_{\mathrm{p}}(t)=\boldsymbol{u}(t)-\hat{\boldsymbol{u}}(t)$ 表示定位误差，且在每个时刻上服从高斯分布[18]，则定位误差的 CRLB 可表示为[19]

$$\mathbb{E}\left\{\left\|\boldsymbol{e}_{\mathrm{p}}(t)\right\|^2\right\}\geqslant\mathrm{Tr}\left(\boldsymbol{J}_{\mathrm{u}}^{-1}\left(\boldsymbol{p}_{\mathrm{p}}(t)\right)\right) \quad (7)$$

其中，$\boldsymbol{J}_{\mathrm{u}}^{-1}\left(\boldsymbol{p}_{\mathrm{p}}(t)\right)$ 表示费歇耳信息矩阵（FIM, Fisher information matrix），CRLB 的具体推导过程见附录 1。

1.2 反馈子帧

在反馈子帧中，基站根据反馈得到的用户估计位置 $\hat{\boldsymbol{u}}(t)$ 计算时隙 t 处每个 LED 与移动用户之间的 CSI 估计值。由于存在定位误差 $\boldsymbol{e}_{\mathrm{p}}(t)$，导致 CSI 值也是非理想的。令 $\hat{\boldsymbol{g}}(t)=\left[\hat{g}_1(t),\cdots,\hat{g}_N(t)\right]^{\mathrm{T}}$ 表示 CSI 估计向量，$\Delta\boldsymbol{g}(t)=\left[\Delta g_1(t),\cdots,\Delta g_N(t)\right]^{\mathrm{T}}$ 表示 CSI 估计误差向量，其中 $\hat{g}_i(t)$ 和 $\Delta g_i(t)$ 分别表示时隙 t 处第 i 个 LED 和用户之间的 CSI 估计值和估计误差。根据式(3)，$\hat{g}_i(t)$ 可表示为

$$\hat{g}_i(t)=\frac{\eta\left(\hat{z}_{\mathrm{u}}(t)-z_i\right)^{m+1}}{\left\|\hat{\boldsymbol{u}}(t)-\boldsymbol{v}_i\right\|_2^{m+3}},\forall i\in\mathcal{N} \tag{8}$$

则第 i 个 LED 和用户之间理想的 CSI 可表示为 $g_i(t)=\hat{g}_i(t)+\Delta g_i(t)$。

联立式(3)和式(8)，CSI 误差 $\Delta g_i(t)$ 为

$$\Delta g_i(t)=\frac{\eta\left(z_i-\hat{z}_{\mathrm{u}}(t)\right)^{m+1}}{\left\|\hat{\boldsymbol{u}}(t)-\boldsymbol{v}_i+\boldsymbol{e}_{\mathrm{p}}(t)\right\|_2^{m+3}}-\frac{\eta\left(z_i-\hat{z}_{\mathrm{u}}(t)\right)^{m+1}}{\left\|\hat{\boldsymbol{u}}(t)-\boldsymbol{v}_i\right\|_2^{m+3}} \tag{9}$$

1.3 通信子帧

在通信子帧中，基站根据估计的 CSI 值与用户进行定向通信。具体地，在时隙 t 处，定义 $s_{\mathrm{c}}(t)$ 表示基站发送的通信光信号，满足 $-A\leqslant s_{\mathrm{c}}(t)\leqslant A$、$\mathbb{E}\left\{s_{\mathrm{c}}(t)\right\}=0$ 和 $\mathbb{E}\left\{s_{\mathrm{c}}^2(t)\right\}=\varepsilon$，其中，$A>0$ 和 ε 分别表示通信光信号的峰值和方差，定义 $\boldsymbol{p}_{\mathrm{c}}(t)=\left[\sqrt{P_{\mathrm{c},1}(t)},\cdots,\sqrt{P_{\mathrm{c},N}(t)}\right]^{\mathrm{T}}$ 表示对应的波束成形向量，则发射端通信信号表示为

$$\boldsymbol{x}_{\mathrm{c}}(t)=\boldsymbol{p}_{\mathrm{c}}(t)s_{\mathrm{c}}(t)+\boldsymbol{b} \tag{10}$$

其中，$\boldsymbol{b}=[b,\cdots,b]^{\mathrm{T}}\in R^{N\times 1}$ 表示直流偏置向量。

考虑到 CSI 误差的存在，接收端信号表达式为

$$y_{\mathrm{c}}(t)=\left(\hat{\boldsymbol{g}}_{\mathrm{c}}^{\mathrm{T}}(t)+\Delta\boldsymbol{g}_{\mathrm{c}}^{\mathrm{T}}(t)\right)\boldsymbol{x}_{\mathrm{c}}(t)+n_{\mathrm{c}} \tag{11}$$

其中，n_{c} 表示服从均值为零和方差 σ_{c}^2 的高斯白噪声。

利用 α-β-γ（ABG）分布[5]，移动用户在时隙 t 处的可达通信速率为

$$R_{\mathrm{c}}(t)=\frac{1}{2}\mathrm{lb}\left(1+\frac{\left|\left(\hat{\boldsymbol{g}}^{\mathrm{T}}(t)+\Delta\boldsymbol{g}^{\mathrm{T}}(t)\right)\boldsymbol{p}_{\mathrm{c}}(t)\right|^2\mathrm{e}^{1+2(\alpha+\lambda\varepsilon)}}{2\pi\sigma_{\mathrm{c}}^2}\right) \tag{12}$$

其中，α、β 和 λ 为 ABG 参数。

令 $P_{\mathrm{o}}^{\max}$ 和 P_{e} 分别表示每个 LED 的最大光功率约束和电功率约束，则定位和通信的实际功率约束为[20]

$$\begin{aligned}&0\leqslant P_{\mathrm{p},i}\leqslant\xi,\forall i\in\mathcal{N}\\&0\leqslant P_{\mathrm{c},i}\leqslant\xi,\forall i\in\mathcal{N}\end{aligned} \tag{13}$$

其中，$\xi=\min\left\{\frac{b^2}{A^2},P_{\mathrm{e}}-\frac{b^2}{\varepsilon},\frac{\left(P_{\mathrm{o}}^{\max}-b\right)^2}{A^2}\right\}$。

1.4 问题建模

在满足定位精度要求、总发送功率约束和 LED 实际功率约束条件下，最大化移动用户在整个移动时间 T 上的平均可达速率。数学上，可达速率最大化可建模为如下问题

$$\begin{aligned}&\max_{\left\{\boldsymbol{P}_{\mathrm{p}}(t),\boldsymbol{P}_{\mathrm{c}}(t)\right\}_{t=1}^{T}}\ \mathbb{E}\left\{R_{\mathrm{c}}(t)\right\}\\&\text{s.t. C1: }\sqrt{\mathrm{Tr}\left(\boldsymbol{J}_{\mathrm{u}}^{-1}\left(\boldsymbol{p}_{\mathrm{p}}(t)\right)\right)}\leqslant\chi\\&\quad\text{C2: }\sum_{i=1}^{M}\left(P_{\mathrm{p},i}(t)+P_{\mathrm{c},i}(t)\right)\leqslant P_{\mathrm{total}}\\&\quad\text{C3: }0\leqslant P_{\mathrm{p},i}\leqslant\zeta,0\leqslant P_{\mathrm{p},i}\leqslant\zeta\end{aligned} \tag{14}$$

其中，χ 表示定位误差门限，P_{total} 表示每个时隙上总的发送功率门限。

由于用户具有移动性，问题(14)是在总时隙 T 上的组合优化问题，传统的优化方法需要在大空间上进行搜索，很难以较低的时间复杂度得到此类问题的高质量解[15]。此外，由于目标函数和约束 C1 的影响，问题(14)的优化变量相互耦合，很难直接获得功率分配的解析解。因此，本文提出了一种基于 DRL 的功率分配算法，以高效地解决该问题。

2 算法设计

2.1 强化学习问题建模

作为机器学习的一个重要分支，强化学习旨在通过不断地“试错”与环境交互，进而学习到最佳的策略，以最大化系统的长期奖励或者实现特定的目标。强化学习方法的训练过程可被建模为形如 $(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R})$ 的马尔可夫决策过程[21]，其中，$\mathcal{S}$ 表示状态空间，包含了系统完整的状态信息；$\mathcal{A}$ 表示动作空间，包含了有限个可能采取的动作；$\mathcal{P}$ 表示状态转移概率的集合；$\mathcal{R}$ 表示系统即时奖励的集合。具体地，在时隙 t 处，处于状态 $s_t\in\mathcal{S}$ 的智能体会根据某一策略 $\theta\left(a_t|s_t\right)$ 执行动作 $a_t\in\mathcal{A}$，然后通

过与环境交互获得即时奖励 $r_t \in \mathcal{R}$，接着会根据概率 $p\left(s_{t+1}|s_t,a_t\right)$ 转移到下一状态 $s_{t+1} \in \mathcal{S}$。在本节中，状态、动作和奖励的定义如下。

状态。在 VLPC 一体化系统中，定位功率的分配决定了定位的准确性和 CSI 估计值的准确性。进一步，由于 LED 根据 CSI 估计值进行定向通信，CSI 估计值的准确性在很大程度上影响系统的功率分配策略。因此，在时隙 t 处，定义状态 s_t 为所有 LED 到用户之间的 CSI 估计值，其可以通过上一个时隙 t 处分配的定位功率，利用式(6)计算得

$$s_t = \left\{\hat{g}_1(t),\cdots,\hat{g}_N(t)\right\} \tag{15}$$

动作。当智能体处于状态 s_t 时，会进行定位功率和通信功率的分配，则 a_t 定义为

$$a_t = \left\{\left\{P_{\mathrm{p},1}(t),\cdots,P_{\mathrm{p},N}(t)\right\},\left\{P_{\mathrm{c},1}\left(t\right),\cdots,P_{\mathrm{c},N}(t)\right\}\right\} \tag{16}$$

奖励。在每个训练回合中，智能体会根据当前所处的状态 s_t 选择一个动作 a_t 执行，然后从环境中获得一个奖励值 r_t 作为反馈。由于问题(14)的优化目标是在满足约束的条件下最大化用户的平均可达通信速率，则 r_t 定义为[15]

$$r_t = \begin{cases} R_{\mathrm{c}}(t), & \text{当满足C1、C2和C3时} \\ 0, & \text{其他} \end{cases} \tag{17}$$

通过不断地与环境交互，智能体可以学习到一个最优策略 θ^*，以最大化长期折扣奖励，定义为

$$R_t^{\gamma} = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \tag{18}$$

其中，$\gamma \in [0,1]$ 表示折扣系数，用于智能体权衡当前奖励和未来奖励的重要性。当 γ 越靠近 1 时，表示智能体越重视长期奖励；当 γ 越靠近 0 时，表示智能体仅注重短期奖励。

强化学习的目标是找到一个策略以最大化智能体获得的长期折扣奖励。令 $Q_\theta\left(s_t,a_t\right)$ 表示 Q 值函数，用于评价智能体在策略 θ 的指导下，以状态 s_t 选择动作 a_t 的价值，可根据贝尔曼方程推导如下[22]

$$Q_\theta\left(s_t,a_t\right) = \mathbb{E}\left[r_t + \gamma \max_{a_{t+1}} Q_\theta\left(s_{t+1},a_{t+1}\right)\middle|s_t,a_t\right] \tag{19}$$

2.2　基于 DDPG 的功率分配算法

问题(14)可以通过使用基于强化学习的 DQN 算法来解决，其中关键步骤是将功率在可行域内量化为一些离散值。然而由于量化误差难以避免，这可能导致某些关键的功率分配取值丢失。虽然可以通过增大量化等级减少误差，但同时也会增大 DQN 的搜索空间，给算法收敛带来困难。受文献[24]启发，本文提出了一种基于深度确定性策略梯度的功率分配算法，其框架如图 3 所示。

基于 Actor-Critic 模式，DDPG 网络由 4 个深度神经网络（DNN, deep neural network）组成，包括一个权值为 ϕ 的 Actor 网络 $\mu(s_i;\phi)$，用于输出对应的动作；一个权值为 ρ 的 Critic 网络 $Q(s_i,a_i;\rho)$，用于评估所选择动作的 Q 值；一个权值为 ϕ' 的目标 Actor 网络 $\mu(s_i;\phi')$；一个权值为 ρ' 的目标 Critic 网络 $Q(s_{i+1},\mu(s_{i+1};\phi');\rho')$，用于产生训练的 Q 值。

为了保证问题(14)中光功率约束 C3，需要对 DDPG 网络的动作输出层进行改写。具体地，令

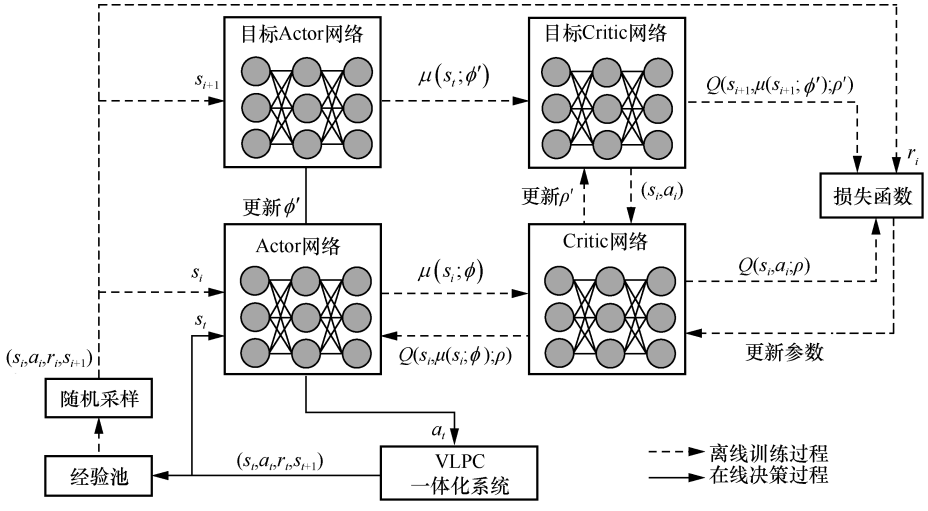


图 3　DDPG 算法框架

$x_t=\mu(s_i;\phi)$ 表示 Actor 网络的输出，使用高斯随机噪声 n 平衡对新动作的探索和对已知动作的利用。考虑到定位和通信功率的上界 ξ，DDPG 网络的实际动作 a_t 可表示为

$$a_t=\left[\left(S\left(x_t\right)+n\right)\zeta\right]_0^{\zeta} \tag{20}$$

其中，$S\left(x_t\right)=\dfrac{1}{1+\mathrm{e}^{-x_t}}$ 为 sigmoid 函数。式(20)确保每一个 LED 的发送功率被约束到 $[0,\zeta]$，以使其满足问题(14)中的约束条件 C3，若为其他激活函数则不满足问题(14)中的约束条件 C3。

当在线决策处于初始阶段时，基站作为智能体，先给所有 LED 分配相等的定位功率，通过式(6)获取初始的用户估计位置。在时隙 t 处，基于式(8)，智能体将估计的 CSI 视为状态 s_t，并将其发送到 DDPG 单元中，然后基于式(20)，得到对应的功率分配动作 a_t，并利用 a_t 中分配的通信功率获得用户的即时通信速率，进而基于式(17)获得奖励 r_t。之后，随着用户移动到下一位置，通过 a_t 中分配的定位功率得到用户移动后的估计位置，进而获得状态 s_{t+1}。

本节采用随机从经验池 $\mathcal{D}$ 中采样 Z 个样本 (s_i,a_i,r_i,s_{i+1}) 的方法，以打破训练时数据的相关性。具体而言，在离线训练过程中，利用目标 Actor 网络 $\mu(s_t;\phi')$ 和目标Critic网络 $Q(s_{i+1},\mu(s_{i+1};\phi');\rho')$ 生成用于训练的目标 Q 值，即

$$y_i=r_i+\gamma Q(s_{i+1},\mu(s_{i+1};\phi');\rho') \tag{21}$$

同时，Critic 网络通过最小化均方误差（MSE, mean square error）损失函数来更新其权重 ρ，MSE 函数曲线光滑、连续、处处可导，便于使用梯度下降算法，是一种常用的损失函数。而且随着误差的减小，梯度也在减小，这有利于收敛，即使使用固定的学习速率，也能较快地收敛到最小值。因此本文选用 MSE 作为误差度量函数，即

$$L(\rho)=\frac{1}{Z}\sum_{i=1}^{M}\left(y_i-Q\left(s_i,a_i;\rho\right)\right)^2 \tag{22}$$

根据确定性策略梯度定理[23]，Actor 网络 $\mu\left(s_t;\phi\right)$ 在获得更大的累积折扣奖励的方向上更新其权重 ϕ，即

$$\nabla_{\phi}J(\phi)=\mathbb{E}\left[\nabla_{\phi}\mu\left(s_i;\phi\right)\nabla_{a_i}Q\left(s_i,a_i;\rho\right)\Big|_{a_i=\mu(s_i;\phi)}\right] \tag{23}$$

其中，$J(\phi)=\mathbb{E}\left[Q\left(s_i,a_i;\rho\right)\Big|_{a=\mu(s_i;\phi)}\right]$ 表示在所有状态都遵循策略 θ 的预期总回报。

使用 $\mathcal{D}$ 中的 Z 个随机采样元组，式(23)可以通过近似计算式(24)得到

$$\nabla_{\phi}J(\phi)\approx\frac{1}{Z}\sum_{i=1}^{Z}\left[\nabla_{\phi}\mu\left(s_i;\phi\right)\nabla_{a_i}Q\left(s_i,\mu\left(s_i;\phi\right);\rho\right)\right] \tag{24}$$

最后，使用软更新的方式更新目标网络的权值

$$\begin{aligned}\rho'&\leftarrow\tau\rho+(1-\tau)\rho'\\ \phi'&\leftarrow\tau\phi+(1-\tau)\phi'\end{aligned} \tag{25}$$

其中，τ 表示软更新系数，且满足 $0\leqslant\tau\leqslant1$。算法 1 总结了基于 DDPG 的功率分配算法。

算法 1 基于 DDPG 的功率分配算法

初始化 经验池 $\mathcal{D}$，批次 Z，回合数 E，每回合训练总时隙 T，折扣因子 γ，软更新系数 τ，DDPG 各个网络的权值 ϕ、ϕ'、ρ 和 ρ'

1) for $e=1,2,\cdots,E$, 进入循环
2) 收到初始状态 s_1；
3) for $t=1,2,\cdots,T$，进入循环
4) 基于式(20)选择动作 a_t；
5) 基于式(17)获得奖励 r_t，用户移动到下个位置，获得下一时隙CSI估计值作为 s_{t+1}；
6) 将 $\left(s_t,a_t,r_t,s_{t+1}\right)$ 存储在 $\mathcal{D}$ 中；
7) 分别通过式(22)和式(24)更新 ρ 和 ϕ；
8) 通过式(25)更新 ρ' 和 ϕ'；
9) end for；
10) end for；
11) 输出 ϕ、ϕ'、ρ 和 ρ'

3 仿真分析

为了验证本文算法的有效性，本节给出了数值结果用于评估所提出的基于 DDPG 功率分配算法的性能，并与 DQN 算法和等功率分配算法进行对比。在仿真中，考虑一个部署在 5 m×5 m×3 m 房间内的 VLCP 一体化系统，其中距离单位为 m。将房间建模为三维坐标系 (X,Y,Z)，房间的一角为坐标原点(0,0,0)，VLPC 系统参数如表 1 所示，DDPG 算法参数如表 2 所示。基站包括 4 个 LED，其坐标分别为(1,1,3)，(3,3,3)，(1,3,3)，(3,1,3)。移动用户起点坐标为(2,2,1.3)，移动范围为半径 2 m 的圆形区域，移动速度为 0.2 m/s，且在每个时隙上从向前、向后、向左与向右这 4 种移动方向中随机选择一种。

表 1　　VLPC 系统参数

参数	参数值
LED 半功率角 $\psi_{\frac{1}{2}}$	60°
接收机视场角 φ_{FOV}	90°
折射率 n_r	1.5
接收机感光面积 A_r /cm^2	1.0
聚光器增益 $\Gamma(\varphi)$ /(A·W^{-1})	1.0
滤光片增益 F_s (A·W^{-1})	1.0
噪声功率谱密度 $\sigma_{\mathrm{p}}^2,\sigma_{\mathrm{c}}^2$ /dBm	−98.82
信号幅度 A/V	1.0
信号方差 ε	1.0
直流偏置 b /A	0.447 2
平均电功率 P_{e} /W	1.0
最大光功率 $P_{\mathrm{o}}^{\max}$ /W	2.236
总功率门限 P_{total} /W	3

表 2　　DDPG 算法参数

参数	参数值
回合数 E	1 000
每回合总时隙 T	20
经验池 $\mathcal{D}$ 大小	3 000
采样批次 Z	32
优化器	Adam
折扣因子 γ	0.9
Dropout	0.1
Actor 网络学习率	0.000 2
Critic 网络学习率	0.000 4
软更新系数 τ	0.000 01

为了直观地比较 DDPG 和 DQN 方案在平均可达速率上的差距，图 4 给出了 2 种方案的平均可达速率随量化等级的变化情况，以 DDPG 方案为基准，其中量化等级集合设置为 $\{3,6,10,15,20,30,40,50\}$。从图 4 可以看到，当量化等级从 3 提高到 50 时，DQN 方案的平均可达速率先逐渐增加，然后开始减小。这说明通过增大量化等级可以提升 DQN 方案的性能，但是过大的动作空间会导致 DQN 方案的实际训练困难，并且通过简单地增加动作空间的维度来消除量化误差是不可行的。而 DDPG 方案本质上不需要对功率进行量化取值，因而其性能优于 DQN 方案。由图 4 可知，不同量化等级下的 DDPG 方案的平均可达速率均大于 DQN 方案。

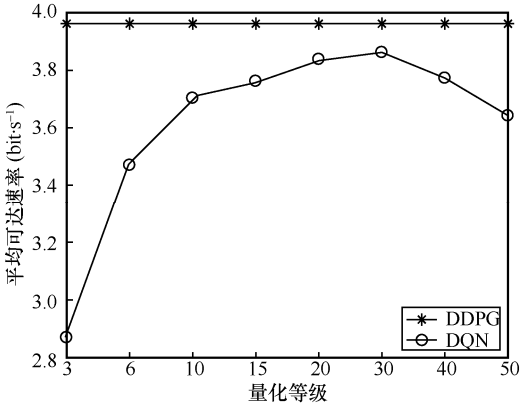


图 4　平均可达速率随量化等级的变化情况

图 5 给出了 3 种方案的平均可达速率随训练回合数的变化曲线以及 2 种量化等级下的 DQN 变化曲线。从图 5 中可以看出，DDPG 和 DQN 方案的平均可达速率刚开始都比较小，经过训练后都分别收敛到一个相对稳定的值，这说明这 2 种基于 DRL 的方案均能在与环境的不断交互中学习到新信息，具备良好的收敛性，而等功率分配方案通过在每个时隙上给 LED 分配相等的定位和通信功率，并没有自主决策的能力，故难以取得较好的通信速率。由图 5 可知，量化等级 k=30 的 DQN 方案的平均可达速率优于量化等级 k=10 的 DQN 方案的平均可达速率，并由图 4 知当量化等级 k>30 时，平均可达速率随之减少。由图 5 可知，DQN 方案的收敛训练回合数为 200，DDPG 为 300。本文采用分布式 DQN[15,24]方案，使用多 DQN 单元分布式运行结构以减少动作空间维度，加快了收敛速度。此外还可以看出，当所提出的 2 种方案收敛后，基于 DDPG 的方案在平均可达速率上优于 DQN 方案。这是因为 DDPG 单元采用 Actor-Critic 架构来构造策略函数来直接输出所选择的动作，可以解决 DQN 因量化功率取值导致的误差问题，因而能取得更优的性能。

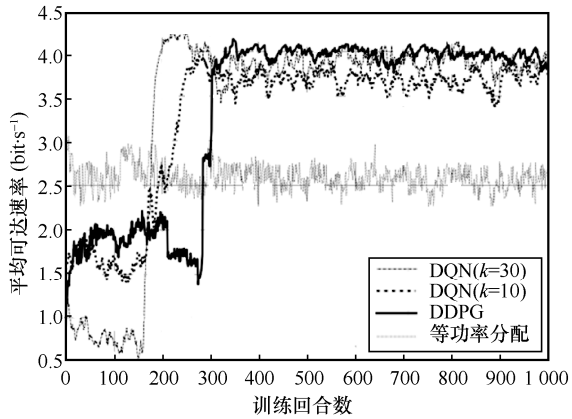


图 5　平均可达速率随训练回合数的变化曲线

图 6 给出了 3 种功率分配方案的定位误差的 $\sqrt{\text{CRLB}}$ 的累积分布函数（CDF, cumulative distribution function）曲线，其中 $\sqrt{\text{CRLB}}$ 门限值设置为 $\chi = 0.2$ m。从图 6 可以看出，DDPG 和 DQN 方案均有 90%以上的概率满足给定的 $\sqrt{\text{CRLB}}$ 门限值，其中，基于 DDPG 的方案满足门限值的概率约为 97%，基于 DQN 的方案满足门限值的概率约为 92%，而等功率分配方案满足门限值的概率仅约为 25%，这说明了基于 DRL 的方案可以有效缓解定位误差带来的影响，并且结合图 5 的结论可知，基于 DDPG 的功率分配方案能取得均优于 DQN 的性能。基于 DQN 的方案对于连续功率变量进行了量化取值，但量化误差不可避免，进而可能导致某些关键的功率分配取值丢失。虽然可以通过增大量化等级减少误差，但同时也会增大 DQN 的搜索空间，给算法收敛带来困难；不同于 DQN 通过概率输出离散的动作，DDPG 根据策略直接生成确定性动作，其取值包含了所有的动作空间，从而有效解决动作空间的维度问题。

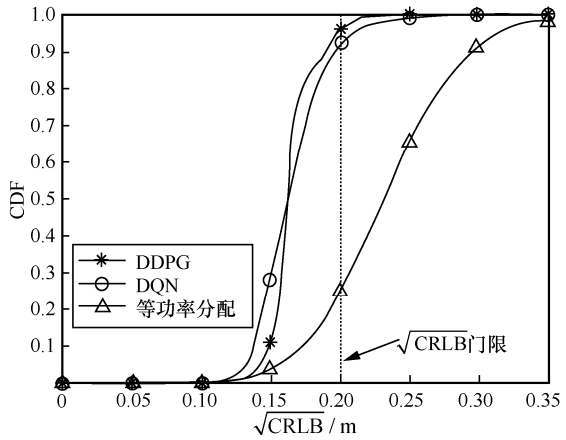


图 6　定位误差的 $\sqrt{\text{CRLB}}$ 的 CDF 曲线

本文通过运行 1 000 次回合所需时间来比较方案的复杂度，等功率分配方案为 1 160 s，DQN 方案为 1 100 s，DDPG 方案为 6 428 s，虽然 DDPG 方案所需时间较长，但 DDPG 方案的平均可达速率最高。

图 7 给出了 3 种方案的平均可达速率随总功率门限 P_{total} 的变化曲线。由式(12)可知，平均可达速率随着功率增加而增加。从图 7 可以看出，3 种方案的平均可达速率都随着 P_{total} 的增加而增加，这是因为随着总功率门限的增加，用于定位的功率就越多，CSI 估计就越准确，且 LED 获得的通信功率也会随之增加，从而使用户的平均可达速率增加。

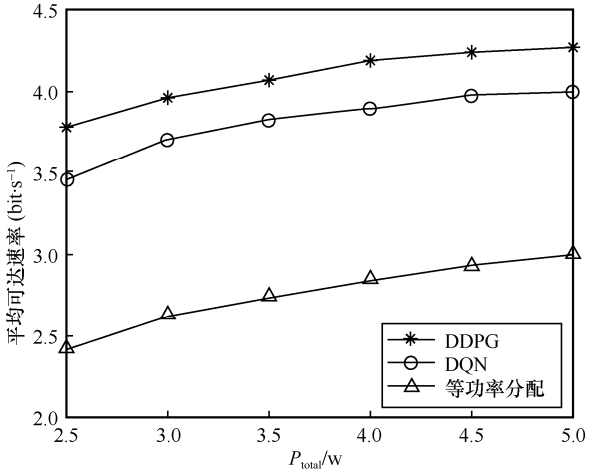


图 7　平均可达速率随总功率门限 P_{total} 的变化曲线

4　结束语

本文提出了一种基于深度强化学习的 VLPC 一体化系统的功率分配方案。首先，提出了定位通信一体化帧结构设计；然后，利用定位信息实现了信道状态信息的估计，并推导了定位误差的 CRLB 和通信速率，阐明了定位精度和通信速率的内在耦合关系；在此基础上，研究了满足 CLRB 门限、LED 实际功率约束下的动态功率分配问题，以最大化移动用户的平均通信速率。由于传统优化方法难以解决该动态功率分配问题，本文提出了基于 DDPG 的 VLPC 动态功率分配方案。仿真结果表明，所提方案能取得良好的通信性能，并能有效缓解定位误差带来的影响。

附录 1　CRLB 的推导

简洁起见，本节推导省略时隙 t。定义用户的三维位置坐标 $\boldsymbol{u} = [x_{\text{u}}, y_{\text{u}}, z_{\text{u}}]^{\text{T}}$ 表示待估计的用户位置向量，根据式(4)，定位信号 $y_{\text{p},i}$ 的似然函数可表示为

$$f\left(y_{\text{p},i};\boldsymbol{u}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\text{p}}} \text{e}^{-\frac{\left(y_{\text{p},i} - g_i x_{\text{p},i}\right)^2}{2\sigma_{\text{p}}}} \tag{26}$$

其对数似然函数可表示为

$$\Lambda(\boldsymbol{u}) = \kappa - \frac{1}{2\sigma_{\text{p}}^2} \sum_{i=1}^{N} \int_0^{T_{\text{p}}} \left(y_{\text{p},i} - g_i x_{\text{p},i}\right)^2 \text{d}t \tag{27}$$

其中，κ 是与未知参数无关的常数，T_{p} 是定位信号的持续时间。

FIM 矩阵 $\boldsymbol{J}_{\text{u}}(\boldsymbol{p}_{\text{p}})$ 是定位功率向量 $\boldsymbol{p}_{\text{p}} = \left[\sqrt{P_{\text{p},1}}, \cdots, \sqrt{P_{\text{p},N}}\right]^{\text{T}}$ 的函数，可表示为

$$\boldsymbol{J}_{\mathrm{u}}(\boldsymbol{p}_{\mathrm{p}})=\begin{bmatrix}-\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial x_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial y_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial z_{\mathrm{u}}}\right)\\ -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial x_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial y_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial z_{\mathrm{u}}}\right)\\ -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}x_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}\partial y_{\mathrm{u}}}\right) & -\mathbb{E}\left(\dfrac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}\partial z_{\mathrm{u}}}\right)\end{bmatrix} \tag{28}$$

其中，一阶导数计算过程为

$$\begin{aligned}\frac{\partial\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p,i}}^2g_i-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial x_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p,i}}^2g_i-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial y_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p,i}}^2g_i-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial z_{\mathrm{u}}}\right)\mathrm{d}t\end{aligned} \tag{29}$$

基于式(29)，二阶导数计算过程为

$$\begin{aligned}\frac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial x_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial x_{\mathrm{u}}}\frac{\partial g_i}{\partial x_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial y_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial y_{\mathrm{u}}}\frac{\partial g_i}{\partial y_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}\partial z_{\mathrm{u}}}&=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial z_{\mathrm{u}}}\frac{\partial g_i}{\partial z_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial y_{\mathrm{u}}}&=\frac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial x_{\mathrm{u}}}=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial x_{\mathrm{u}}}\frac{\partial g_i}{\partial y_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial^2\Lambda(\boldsymbol{u})}{\partial x_{\mathrm{u}}\partial z_{\mathrm{u}}}&=\frac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}\partial x_{\mathrm{u}}}=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial x_{\mathrm{u}}}\frac{\partial g_i}{\partial z_{\mathrm{u}}}\right)\mathrm{d}t\\ \frac{\partial^2\Lambda(\boldsymbol{u})}{\partial y_{\mathrm{u}}\partial z_{\mathrm{u}}}&=\frac{\partial^2\Lambda(\boldsymbol{u})}{\partial z_{\mathrm{u}}\partial y_{\mathrm{u}}}=-\frac{1}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\int_0^{T_{\mathrm{p}}}\left(\left(x_{\mathrm{p},i}^2+g_ix_{\mathrm{p},i}^2-y_{\mathrm{p},i}x_{\mathrm{p},i}\right)\frac{\partial g_i}{\partial y_{\mathrm{u}}}\frac{\partial g_i}{\partial z_{\mathrm{u}}}\right)\mathrm{d}t\end{aligned} \tag{30}$$

由于定位信号满足 $\mathbb{E}\{s_{\mathrm{p},i}\}=0$，$\mathbb{E}\{s_{\mathrm{p},i}^2\}=\varepsilon$，式(28)可重新表示为

$$\boldsymbol{J}_{\mathrm{u}}(\boldsymbol{p}_{\mathrm{p}})=\frac{T_{\mathrm{p}}}{\sigma_{\mathrm{p}}^2}\sum_{i=1}^{N}\left(P_{\mathrm{p},i}\varepsilon+b^2\right)\boldsymbol{\Phi} \tag{31}$$

其中，矩阵 $\boldsymbol{\Phi}$ 可表示为

$$\boldsymbol{\Phi}=\begin{bmatrix}\sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\\ \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\\ \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\dfrac{\partial g_i}{\partial x_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\dfrac{\partial g_i}{\partial y_{\mathrm{u}}} & \sum\limits_{i=1}^{N}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\dfrac{\partial g_i}{\partial z_{\mathrm{u}}}\end{bmatrix} \tag{32}$$

其中，相关微分项可表示为

$$\begin{aligned}\frac{\partial g_i}{\partial x_{\mathrm{u}}}&=-(m+3)\eta\left(z_i-z_{\mathrm{u}}\right)^{-m-1}\left(x_{\mathrm{u}}-x_i\right)\left\|\boldsymbol{u}-\boldsymbol{v}_i\right\|_2^{-m-5}\\ \frac{\partial g_i}{\partial y_{\mathrm{u}}}&=-(m+3)\eta\left(z_i-z_{\mathrm{u}}\right)^{-m-1}\left(y_{\mathrm{u}}-y_i\right)\left\|\boldsymbol{u}-\boldsymbol{v}_i\right\|_2^{-m-5}\\ \frac{\partial g_i}{\partial z_{\mathrm{u}}}&=-(m+1)\eta\left(z_i-z_{\mathrm{u}}\right)^{m}\left\|\boldsymbol{u}-\boldsymbol{v}_i\right\|_2^{-m-5}+\\ &\quad(m+3)\eta\left(z_i-z_{\mathrm{u}}\right)^{m+2}\left\|\boldsymbol{u}-\boldsymbol{v}_i\right\|_2^{-m-5}\end{aligned} \tag{33}$$

根据 CRLB 对任何无偏估计量的均方误差的定义，定位误差 $\boldsymbol{e}_{\mathrm{p}}$ 的 CRLB 可表示为

$$\mathbb{E}\left\{\left\|\boldsymbol{e}_{\mathrm{p}}\right\|^2\right\}\geqslant\mathrm{Tr}\left(\boldsymbol{J}_{\mathrm{u}}^{-1}\left(\boldsymbol{p}_{\mathrm{p}}\right)\right) \tag{34}$$

参考文献：

[1] LIU G Y, HUANG Y H, LI N, et al. Vision, requirements and network architecture of 6G mobile network beyond 2030[J]. China Communications, 2020, 17(9): 92-104.

[2] 迟楠, 陈慧. 高速可见光通信的前沿研究进展[J]. 光电工程, 2020, 47(3): 190687.
CHI N, CHEN H. Progress and prospect of high-speed visible light communication[J]. Opto-Electronic Engineering, 2020, 47(3): 190687.

[3] DAI L L, WANG B C, YUAN Y F, et al. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends[J]. IEEE Communications Magazine, 2015, 53(9): 74-81.

[4] JIANG W, HAN B, HABIBI M A, et al. The road towards 6G: a comprehensive survey[J]. IEEE Open Journal of the Communications Society, 2021, 2: 334-366.

[5] MA S, LI H, HE Y, et al. Capacity bounds and interference management for interference channel in visible light communication networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(1): 182-193.

[6] JIANG R, WANG Q, HAAS H, et al. Joint user association and power allocation for cell-free visible light communication networks[J]. IEEE Journal on Selected Areas in Communications, 2018, 36(1): 136-148.

[7] SHI J Y, ZHU X, WANG F M, et al. Net data rate of 14.6 Gbit/s underwater VLC utilizing silicon substrate common-anode five primary colors LED[C]//Proceedings of Optical Fiber Communication Conference. Washington: OSA, 2019: 1-3.

[8] YANG S H, JUNG E M, HAN S K. Indoor location estimation based on LED visible light communication using multiple optical receivers[J]. IEEE Communications Letters, 2013, 17(9): 1834-1837.

[9] YANG S H, KIM H S, SON Y H, et al. Three-dimensional visible light indoor localization using AOA and RSS with multiple optical receivers[J]. Journal of Lightwave Technology, 2014, 32(14): 2480-2485.

[10] WU Y C, HSU K L, LIU Y, et al. Using linear interpolation to reduce the training samples for regression based visible light positioning system[J]. IEEE Photonics Journal, 2020, 12(2): 1-5.

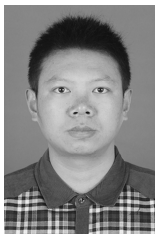
[11] YANG H L, ZHONG W D, CHEN C, et al. QoS-driven optimized design-based integrated visible light communication and positioning for indoor IoT networks[J]. IEEE Internet of Things Journal, 2020, 7(1): 269-283.

[12] LIN B J, TANG X, GHASSEMLOOY Z, et al. Experimental demonstration of an indoor VLC positioning system based on OFDMA[J]. IEEE Photonics Journal, 2017, 9(2): 1-9.

[13] YANG H L, CHEN C, ZHONG W D, et al. Demonstration of a quasi-gapless integrated visible light communication and positioning sys-

tem[J]. IEEE Photonics Technology Letters, 2018, 30(23): 2001-2004.
[14] YANG H L, DU P F, ZHONG W D, et al. Reinforcement learning-based intelligent resource allocation for integrated VLCP systems[J]. IEEE Wireless Communications Letters, 2019, 8(4): 1204-1207.
[15] WANG X M, ZHANG Y H, SHEN R J, et al. DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems[J]. IEEE Internet of Things Journal, 2020, 7(8): 7279-7294.
[16] CHU M, LI H, LIAO X W, et al. Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems[J]. IEEE Internet of Things Journal, 2019, 6(2): 2009-2020.
[17] KAHN J M, BARRY J R. Wireless infrared communications[J]. Proceedings of the IEEE, 1997, 85(2): 265-298.
[18] STEVENS N, STEENDAM H. Magnitude of the distance estimation bias in received signal strength visible light positioning[J]. IEEE Communications Letters, 2018, 22(11): 2250-2253.
[19] KESKIN M F, SEZER A D, GEZICI S. Optimal and robust power allocation for visible light positioning systems under illumination constraints[J]. IEEE Transactions on Communications, 2019, 67(1): 527-542.
[20] MA S, ZHANG F, LI H, et al. Simultaneous lightwave information and power transfer in visible light communication systems[J]. IEEE Transactions on Wireless Communications, 2019, 18(12): 5818-5830.
[21] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.
[22] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[23] XU Y H, YANG C C, HUA M, et al. Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications[J]. IEEE Access, 2020, 8: 18797-18807.
[24] CALABRESE F D, WANG L, GHADIMI E, et al. Learning radio resource management in RANs: framework, opportunities, and challenges[J]. IEEE Communications Magazine, 2018, 56(9): 138-145.

[作者简介]



马帅（1986– ），男，山东日照人，博士，中国矿业大学副教授，主要研究方向为语义通信和可见光通信定位一体化等。

李兵（1997– ），男 ，河南商丘人，中国矿业大学硕士生，主要研究方向为可见光通信和可见光定位。

盛海鸿（1998– ），女，江苏泰州人，中国矿业大学硕士生，主要研究方向为无线通信和可见光通信。

谷荣妍（1999– ），女，安徽六安人，中国矿业大学硕士生，主要研究方向为可见光通信和可见光定位。

周辉（1997– ），男，湖北荆州人，中国矿业大学硕士生，主要研究方向为可见光通信和定位通信一体化。

王洪梅（1983– ），女，山东诸城人，博士，中国矿业大学副教授，主要研究方向为无线通信。

王悦（1994– ），女，吉林四平人，博士，中国矿业大学副教授，主要研究方向为光通信、微波光子学、集成光子器件等。

李世银（1971– ），男，四川犍为人，博士，中国矿业大学教授、博士生导师，主要研究方向为煤矿信息化和移动目标定位。