研究与开发

基于多臂赌博机的RIS辅助MIMO主被动波束成形设计

沈天泽，汪革，宋云超，高天宝，梁汇彬

（南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院，江苏 南京 210003）

摘　要：可重构智能表面（reconfigurable intelligent surface，RIS）因其低功耗、易调节、辅助通信等优势，被广泛应用于毫米波通信领域，现有大多数传输方案利用信道状态信息设计预编码和RIS的被动波束成形矩阵，然而，这将消耗较大的导频开销，导致频谱效率下降。基于此，利用多臂赌博机（multi-armed bandit，MAB）算法进行RIS辅助多输入多输出（multiple-input multiple-output，MIMO）系统的波束成形设计，该算法从历史数据中获取信道协方差矩阵，并用于波束成形设计，以降低导频开销。具体来说，将被动波束成形矩阵设计问题建模为MAB问题，结合线性上置信界（linear upper confidence bound，LinUCB）算法框架来估计信道协方差矩阵，将有效频谱效率设置为奖励、RIS相移向量设置为动作，提出利用层级贪婪搜索算法选择最大化有效频谱效率之和的方法获取相移向量。仿真结果表明，所提出的算法在减少导频开销、提高有效频谱效率方面表现良好，展示了其优越性。

关键词：可重构智能表面；多输入多输出；多臂赌博机；线性上置信界算法

中图分类号：TN928

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2026008

RIS-assisted MIMO active and passive beamforming design based on multi-armed bandit

Shen Tianze, Wang Ge, Song Yunchao, Gao Tianbao, Liang Huibin

College of Electronic and Optical Engineering & College of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: Reconfigurable intelligent surface (RIS) has gained significant attention in millimeter-wave communications due to its advantages, including low power consumption, easy tunability, and enhanced auxiliary communication capabilities. Most existing transmission schemes employ channel state information to design precoding and passive beamforming matrices for RIS. However, this approach incurs substantial pilot overhead, thereby reducing spectral efficiency. To address this issue, a beamforming design scheme for RIS-assisted multiple-input multiple-output (MIMO) systems based on the multi-armed bandit (MAB) algorithm was proposed. The channel covariance matrix

收稿日期：2025−04−04；修回日期：2025−10−09

通信作者：宋云超，songyc@njupt.edu.cn

基金项目：国家自然科学基金面上项目（No.62371249）

Foundation Item: The General Program of the National Natural Science Foundation of China (No.62371249)

was estimated using historical data via the MAB framework, which helped to reduce pilot overhead. Specifically, the passive beamforming matrix design was formulated as an MAB problem and solved using the linear upper confidence bound (LinUCB) algorithm to estimate the channel covariance matrix. The effective spectral efficiency was defined as the reward, while the RIS phase shift vector constitutes the action. The phase shift vector that maximizes the sum of effective spectral efficiency was selected through a hierarchical greedy search algorithm. Simulation results demonstrate that the proposed algorithm effectively reduces pilot overhead and enhances spectral efficiency, thereby confirming its superiority.

Key words: reconfigurable intelligent surface, multiple-input multiple-output, multi-armed bandit, LinUCB algorithm

0 引言

随着信息技术的不断发展，提升通信系统的传输效率、降低传输时延及优化能耗比，成为未来无线通信面临的重大任务[1]。为了提升无线移动通信质量，学者们提出了多项关键技术，包括毫米波[2]、多输入多输出（multiple-input multiple-output，MIMO）[3]、可重构智能表面（reconfigurable intelligent surface，RIS）技术[4-5]等。其中，RIS技术因成本低廉、应用场景灵活，被广泛研究。

RIS，又称智能反射面，是一种由无源/有源射电器件组成的人工平面阵列结构，其中每个无源/有源反射元件都可以通过现场可编程逻辑门阵列（field-programmable gate array，FPGA）等智能控制器单独控制调节[6]。智能器件可以控制反射元件来有效调节入射信号相位与振幅，向特定方向反射通信信号，从而改变无线通信信道环境[7]，提升通信网络性能。由于RIS的被动结构，其功耗极低，成本低廉，且在反射过程中几乎没有额外的热噪声[8-9]，因此，RIS可以在许多场景下大规模架设，且不需要考虑元件与元件之间的干扰。此外，在视距链路被障碍物遮挡时，RIS也能构建出虚拟直射路径连接基站与用户，从而进行稳定通信[10]。因此，RIS已被广泛应用到无线通信领域[11]。RIS辅助下的无线通信成为当前研究的热点。设计RIS的波束成形矩阵也成为一项提高信号质量、减小干扰的信号处理技术。

近年来，RIS辅助通信系统中波束成形的研究工作取得了不少进展。文献[12]通过将优化问题解耦，即将MIMO系统中传统的功率最小化问题和RIS被动波束成形矩阵设计问题分开处理，旨在实现基站发射功率的最小化，并采用半正定松弛（semidefinite relaxation，SDR）优化技术和交替优化算法，提高了系统的性能。文献[13]在不同场景下也做了类似的工作，联合优化基站主动波束成形和RIS被动波束成形，提出一种基于分数规划、最大最小化方法和流形优化方法的高效交替算法，将得到的非凸优化问题转化为两个可解的子问题，从而分别优化RIS相移向量和基站预编码矩阵，并迭代求解。然而，在快速变化的信道环境中，用户需要频繁发送导频信号，以更新瞬时信道状态信息（channel state information，CSI）。该方法往往会占用大量导频，降低了有效信号的发射占比。为了解决这一问题，文献[14]提出了一种包含稀疏矩阵分解阶段和矩阵完成阶段的两阶段算法。在精确估计出级联信道的基础上，该算法不仅能够有效减少所需导频的数量，还能够充分利用信道的低秩性和稀疏性特征，提升信道估计的性能。为了进一步减小导频开销，降低复杂度，相关研究人员借助统计CSI，设计相移最大化遍历频谱效率。文献[15]提出了一种新的双时间尺度传输协议，利用统计CSI获取无源RIS相移，再根据获得的RIS相移，在每个时隙上基于瞬时CSI进行预编码设计，大幅降低了开销及算法的复杂度。此外，也有研究采用

了双时间尺度的思想，以降低导频开销，即利用瞬时 CSI 设计预编码，利用统计 CSI 进行被动波束成形矩阵设计。文献[16]将粒子群优化算法引入 RIS 被动波束成形矩阵设计中，通过减少信道样本来优化大时间尺度上 RIS 的相移，基于历史的瞬时 CSI 用注水法进行功率分配。总体而言，在大规模 MIMO 通信系统中，瞬时 CSI 变化频繁，但是统计 CSI 变化较为缓慢，只有当基站、RIS 和用户之间的相对位置发生变化时才会改变。因此，在 RIS 辅助的大规模 MIMO 系统场景下，利用统计 CSI 进行被动波束成形矩阵设计能够减小获得瞬时 CSI 所需的导频开销[17]。

此外，强化学习因处理优化问题的强大能力受到越来越多人的关注，也被运用于 MIMO 系统。强化学习通过与环境的交互来学习最优策略，依靠探索（未知领域）和利用（现有信息）的方式来获得反馈，以最大化累积奖励，即使在不同的环境下也能通过不断的交互，获得最优的结果[18]。其中，多臂赌博机（multi-armed bandit，MAB）作为强化学习特例，是一种无监督的在线学习算法，不需要像机器学习与深度学习一样事先进行标签训练。在通信系统中，为了降低导频开销，可将 MAB 运用到 MIMO 系统主动和被动波束成形矩阵设计中。基于 MAB 的 RIS 被动波束成形矩阵设计不需要瞬时 CSI，同时避免了对统计 CSI 的获取，因此，降低了训练开销。文献[19]将 MAB 运用到单用户上行通信场景中，将相移设计转化为上下文 MAB 问题，利用历史通信数据，用流形优化方法获得信道协方差矩阵，以估计相移。

鉴于上述分析，本文针对 RIS 辅助多用户 MIMO 系统的下行通信链路，设计出一种基于 MAB 方法的传输方案，以最大化有效频谱效率。该方案能够利用 MAB 从历史数据中获取信道，并选择相移，有效降低了导频开销和用户间干扰，提高了通信系统性能。本文的主要贡献如下。

（1）将被动波束成形矩阵设计问题建模成 MAB 问题。具体而言，设臂为 RIS 相移向量，该向量属于一个高维连续空间。利用线性上置信界（linear upper confidence bound，LinUCB）框架，可推导出各用户信道能量之和与信道协方差矩阵之间的线性关系。将历史时隙内接收到的信号及相移向量设定为上下文信息，在每个时隙内，由于信道能量与信道协方差矩阵间存在线性关系，因此，可以根据收集到的上下文信息对级联信道协方差矩阵进行估计。

（2）提出层级贪婪搜索算法来设计被动波束成形矩阵，以最大化有效频谱效率。层级贪婪搜索算法包括 3 层：最优贪婪法、线性估计补充法、随机收缩探索法。具体而言，最优贪婪法以较高概率贪婪地选择能够最大化有效频谱效率的臂；线性估计补充法以一定概率基于估计的协方差矩阵设计 RIS 相移向量，并作为次优臂补充至臂集合中；最后，在部分时隙内采用随机收缩探索法在最优臂附近进行探索。

1　系统模型

RIS 辅助的多用户 MIMO 系统模型如图 1 所示，其中，基站通过 RIS 同时向 K 个用户终端发送信息。基站和 RIS 系统都配备了 $M=M_x\times M_y$ 和 $N=N_x\times N_y$ 的均匀平面阵列（uniform planar array，UPA），M_x 和 M_y 分别为基站在水平和竖直方向上的天线阵列个数，N_x 和 N_y 分别为 RIS 辅助系统在水平方向和竖直方向上的无源反射元件个数。集合 $\{1,2,\cdots,k,\cdots,K\}$ 表示用户集，且每个用户终端都单独配备一根天线。本文采用 Saleh Valenzuela 几何信道模型对毫米波信道进行建模[20]。

基站处的响应向量为：

$$\boldsymbol{a}_M\left(\psi_l^{\mathrm{v}},\psi_l^{\mathrm{h}}\right)=\frac{1}{\sqrt{M}}\left[1,\cdots,\mathrm{e}^{\mathrm{j}\pi((M_x-1)\sin\psi_l^{\mathrm{v}}\sin\psi_l^{\mathrm{h}}+(M_y-1)\cos\psi_l^{\mathrm{h}})}\right]^{\mathrm{T}}\quad(1)$$

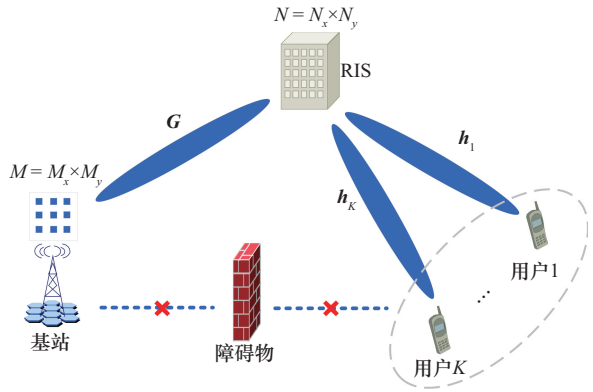


图1　RIS辅助的多用户MIMO系统模型

RIS处的响应向量$\boldsymbol{a}_N\left(\gamma_l^{\mathrm{v}},\gamma_l^{\mathrm{h}}\right)$与基站处的响应向量类似。本文用$\boldsymbol{G}(t)\in\mathbf{C}^{N\times M}$表示基站到RIS系统$t$时隙的信道，信道$\boldsymbol{G}(t)$可表示为：

$$\boldsymbol{G}(t)=\sqrt{\frac{MN}{L}}\sum_{l=1}^{L}\alpha_l\boldsymbol{a}_N\left(\gamma_l^{\mathrm{v}},\gamma_l^{\mathrm{h}}\right)\boldsymbol{a}_M^{\mathrm{H}}\left(\psi_l^{\mathrm{v}},\psi_l^{\mathrm{h}}\right)\tag{2}$$

其中，L表示信道路径数，$\alpha_l\in\mathcal{CN}(0,1)$表示第$l$条路径上的信道增益，$\psi_l^{\mathrm{h}}$、$\psi_l^{\mathrm{v}}$表示基站天线发射的俯仰角和方位角，$\gamma_l^{\mathrm{h}}$、$\gamma_l^{\mathrm{v}}$表示到达RIS的俯仰角和方位角。

类似地，本文用$\boldsymbol{H}(t)\in\mathbf{C}^{N\times K}$表示RIS到用户终端$t$时隙的信道，其中，$\boldsymbol{H}(t)$由图1中RIS与各用户$k$之间的信道$\boldsymbol{h}_k(t)$组成，表示为：

$$\boldsymbol{H}(t)=[\boldsymbol{h}_1(t),\boldsymbol{h}_2(t),\cdots,\boldsymbol{h}_K(t)]\tag{3}$$

$$\boldsymbol{h}_k(t)=\sqrt{\frac{N}{P}}\sum_{p=1}^{P}\beta_p\boldsymbol{a}_N\left(\kappa_p^{\mathrm{v}},\kappa_p^{\mathrm{h}}\right),k\in K\tag{4}$$

其中，P表示路径数，$\beta_p\in\mathcal{CN}(0,1)$表示第$p$条路径上的增益，$\kappa_p^{\mathrm{h}}$和$\kappa_p^{\mathrm{v}}$分别表示RIS出发的俯仰角和方位角，$\boldsymbol{\theta}(t)\in\mathbf{C}^{N\times 1}$表示$t$时隙RIS相移向量，向量中RIS上第$n$个元素的相移为$\boldsymbol{\theta}_n(t)=\mathrm{e}^{\mathrm{j}\phi_t(n)}$，$\phi_t(n)$是一个在$[0,2\pi)$上的变量。因此，在$t$时隙，基站与用户$k$之间的级联信道$\bar{\boldsymbol{h}}_k(t)$可以表示为：

$$\bar{\boldsymbol{H}}(t)=[\bar{\boldsymbol{h}}_1;\bar{\boldsymbol{h}}_2;\cdots;\bar{\boldsymbol{h}}_k;\cdots;\bar{\boldsymbol{h}}_K]\tag{5}$$

$$\bar{\boldsymbol{h}}_k(t)=\boldsymbol{h}_k^{\mathrm{H}}(t)\,\mathrm{diag}(\boldsymbol{\theta}_t)\boldsymbol{G}(t)\tag{6}$$

本文假设用户终端与基站间存在障碍物或存在较高的路径损耗，因此，对基站到用户的直接链路不作考虑。用户k接收到的信号表示为：

$$y_k(t)=\bar{\boldsymbol{h}}_k(t)\sum_{k=1}^{K}\boldsymbol{w}_k s_k(t)+\boldsymbol{n}_k(t)\tag{7}$$

其中，$\boldsymbol{n}_k(t)\sim\mathcal{CN}(0,\sigma^2)$表示$t$时隙的白高斯噪声，$\boldsymbol{w}_k\in\mathbf{C}^{M\times 1}$表示波束成形向量，$s_k(t)$表示用户发送的信息。

信干噪比（signal to interference plus noise ratio，SINR）和有效频谱效率（effective spectral efficiency，ESE）的具体表达式为：

$$\mathrm{SINR}_{t,k}=\left|\bar{\boldsymbol{h}}_k\boldsymbol{w}_k\right|^2\Big/\left(\sum_{j=1,j\neq k}^{K}\left|\bar{\boldsymbol{h}}_k\boldsymbol{w}_k\right|^2+\sigma^2\right)\tag{8}$$

$$\mathrm{ESE}(t)=\left(1-\frac{\tau}{S}\right)\sum_{k=1}^{K}\mathrm{lb}(1+\mathrm{SINR}_{t,k})\tag{9}$$

其中，τ为导频开销，S为一个相干时间内可用的总符号数。

2　问题表述

考虑一个大时间尺度的场景，其中包含多个相干时间，每个相干时间由T_s个时隙组成，整个大时间尺度包含T个时隙。假设在这些相干时间内，统计CSI保持不变，而在每个相干时间内，瞬时CSI也保持不变，通信帧结构示意图如图2所示。

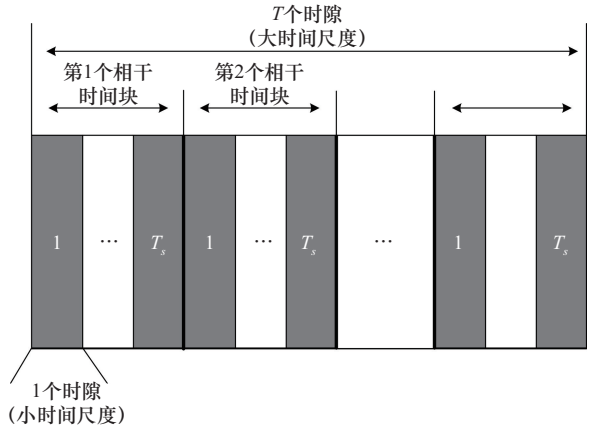


图2　通信帧结构示意图

在小时间尺度场景下，当获得用于通信的RIS相移向量$\boldsymbol{\theta}$时，通过估计瞬时CSI，设计预编码矩阵以抑制用户间干扰。在每个时隙内，基站通过RIS向用户发送长度为$M\times M$的正交导频序列$\boldsymbol{X}$[21]，用户将接收到的信号反馈给基站，此时基站可以根据获得的信号估计出一个包含估计误差的级联信道$\hat{\boldsymbol{H}}$。在发送导频期间，用户接收到的信号为：

$$\boldsymbol{Y}=\bar{\boldsymbol{H}}\boldsymbol{X}^{\mathrm{H}}+\boldsymbol{Z} \tag{10}$$

其中，$\boldsymbol{Z}$表示加性白高斯噪声。当用户获得接收信号后，将信息反馈给基站进行处理，估计的信道便可表示为：

$$\hat{\boldsymbol{H}}=\boldsymbol{Y}\boldsymbol{X} \tag{11}$$

为了减少用户间干扰，利用迫零预编码[22]对预编码矩阵进行设计，即：

$$\boldsymbol{W}=\hat{\boldsymbol{H}}\times\left(\hat{\boldsymbol{H}}\times\hat{\boldsymbol{H}}\right)^{-1} \tag{12}$$

在大时间尺度场景下，假设统计CSI在较大的一个时间段内保持不变。本文考虑在统计CSI不变的时间内，设计RIS上被动波束成形矩阵来最大化若干个时隙内各用户的有效频谱效率之和，具体为：

$$\begin{aligned}\mathrm{P}(1):\quad &\max_{\boldsymbol{\theta}_t}\ \sum_{t=1}^{T}\sum_{k=1}^{K}R_{t,k}\\ &\text{s.t.}\ \ \left|\boldsymbol{\theta}_{i,t}\right|=1, i=1,2,\cdots,N\end{aligned} \tag{13}$$

为了求解P(1)，传统方法在设计RIS被动波束成形矩阵前，需要消耗大量导频信号来获取统计CSI，从而降低了整个时间尺度上的有效频谱效率。为了降低估计统计CSI所需的大量导频，引入MAB方法从历史数据中获取统计CSI，并进行相移设计。由于MAB可以在在线学习的过程中利用历史数据来学习，因此，可以减少估计统计CSI所带来的时间开销和导频开销。

3　层级贪婪搜索算法设计

本文将P(1)建模成MAB问题，其中RIS相移向量$\boldsymbol{\theta}$作为臂，有效频谱效率ESE作为奖励。在每个时隙内，智能体根据环境状态变化调整RIS相移向量，从而影响有效频谱效率并提高通信质量。具体来说，MAB问题的基本构成如下。

（1）智能体。智能体指的是FPGA等智能控制器件，这些设备负责实时监测环境状态变化，并根据反馈调整RIS相移向量。智能体的目标是通过感知环境并选择臂，以最大化累积的有效频谱效率之和。

（2）臂。定义RIS上的n维相移向量$\boldsymbol{\theta}$为臂a，它是一个连续且高维的向量，维度与RIS元件数有关。定义n维向量所处的高维空间为臂空间。为了便于利用MAB算法训练出符合通信要求的相移向量，从该臂空间中离散出臂集A_t，$a\in A_t$。RIS相移向量（臂）的选择将直接影响系统的有效频谱效率。

（3）奖励。定义通过控制器调整RIS相移向量$\boldsymbol{\theta}$（臂a），在基站与各用户间进行通信后获得的实际有效频谱效率之和ESE为选择臂a的奖励。

MAB问题存在探索与利用的困境：探索则意味着尝试不同的臂，以获取更多未知臂的奖励信息；利用则是选择当前已知的最佳臂，以获得最大化的即时奖励。过多的探索会增加选择一些实际奖励并不高的臂的概率，导致短期内奖励减少；过多利用则有可能陷入局部最优解，抑制进一步探索，进而错失选择最优臂的机会。本文所建模的MAB问题中，智能体需要在连续的高维空间中探索和利用，从而优化RIS上的波束成形矩阵。由于高维臂空间是连续的，且臂的数目是无穷的，因此，如何平衡探索和利用的困境，在高维臂空间中探索并最大化平均有效频谱效率，以达到更好的效果，成为本文的研究重点。为此，本文旨在通过设计一种MAB方法来解决探索与利用之间的这一困境。

为了有效解决探索与利用之间的困境，常用的策略包括ε-Greedy策略、上置信界（upper con-

fidence bound，UCB）方法以及汤普森采样等。这些策略旨在有效平衡探索与利用，从而提高整体奖励。其中，ε-Greedy策略适用于解决强化学习问题中MAB固定臂集的情况。然而，本文将RIS上的相移向量作为臂，臂的维度较高，且在臂空间中，臂的分布是连续的。这意味着，直接采用离散方法生成臂集容易错失最优臂，因此，直接运用ε-Greedy策略训练臂集是不可取的。为了在高维臂空间中搜索出最优臂并获得更好的性能，本文基于ε-Greedy策略[23]提出了一种层级贪婪搜索算法来设计被动波束成形矩阵，其包含最优贪婪法、线性估计补充法和随机收缩探索法3种方法，具体内容如下所示：

$$\boldsymbol{\theta}_t=\begin{cases}\boldsymbol{\theta}(\max \mathrm{ESE}), & \varepsilon \geqslant \mathrm{e}^{-\mu t} \\ \boldsymbol{\theta}(\max \left\|\boldsymbol{H}_{\mathrm{eff}}\right\|^2), & \mathrm{e}^{-\omega t} \leqslant \varepsilon<\mathrm{e}^{-\mu t} \\ \exp\left(\mathrm{j} \times 2 \pi \times\left(\boldsymbol{\alpha}_{\max \mathrm{ESE}}+\boldsymbol{n}\right)\right), & \varepsilon<\mathrm{e}^{-\omega t}\end{cases} \tag{14}$$

其中，μ、ω为两个常数，用于调节探索与利用的概率范围，ε为$[0,1]$内的随机数，决定了每个时隙是进行探索还是利用臂集中的臂（RIS相移向量），j为虚数单位，$\boldsymbol{n}$表示随机扰动角度向量，$\boldsymbol{n} \sim \mathcal{CN}(0,10^{-3})$。首先，最优贪婪法以$1-\mathrm{e}^{-\mu t}$的概率，根据对臂的估计值进行贪婪选择[24]，在当前信道状态下最大化有效频谱效率，并提高相干时间内的平均有效频谱效率。其次，鉴于臂空间是高维且连续的，遍历所有臂向量显然不切实际。为了缩小离散范围，提高收敛效率，增加探索到优秀臂的概率，以$\mathrm{e}^{-\mu t}-\mathrm{e}^{-\omega t}$的概率采用线性估计补充法估计信道协方差矩阵，并据此向臂集中补充可能的次优解。最后，运用随机收缩探索法以$\mathrm{e}^{-\omega t}$的概率在最优臂附近进行探索，以避免陷入局部最优解，并进一步提高发现更优秀臂的概率。

因此，设计的层级贪婪搜索算法在ε-Greedy策略的基础上，引入了更为复杂和精细的分层探索与利用机制，并根据时间动态调整探索与利用的侧重，旨在更有效地平衡探索与利用的矛盾，提高寻找最优臂的效率和性能。

3.1 最优贪婪法与随机收缩探索法

在最优贪婪法中，将RIS上的相移$\boldsymbol{\theta}$设置为臂，有效频谱效率之和ESE作为奖励。为了在连续臂空间中应用MAB贪婪策略，在初始化阶段，将随机离散连续高维臂空间，组成初始臂集。t时刻，臂a的平均奖励估计值Q为：

$$Q(a)=\begin{cases}\dfrac{N(a) \times Q(a)+\mathrm{ESE}(t)}{N(a)+1}, & a \text{ 被选择} \\ Q(a), & a \text{ 未被选择}\end{cases} \tag{15}$$

$$N(a)=\begin{cases}N(a)+1, & a \text{ 被选择} \\ N(a), & a \text{ 未被选择}\end{cases} \tag{16}$$

其中，等式右边$N(a)$为前$t-1$时刻内臂a被选择的次数，$Q(a)$为前$t-1$时刻内臂a被选择后的平均奖励。在当前时隙内，当某一臂a被选择后，将结合当前有效频谱效率更新平均奖励值，其被选择次数也相应增加。当其余臂未被选择时，则保持平均奖励与选择次数不变。采用最优贪婪法的每个时隙内，智能体均选择Q值最大的臂，即贪婪地利用当前时隙下平均奖励最大的臂。

此外，为了增强探索并防止陷入局部优解，智能体将会采用随机收缩探索法进行探索。具体而言，智能体以当前时隙所认知的最优相移向量$\boldsymbol{\theta}$作为基准向量，在一定范数距离内更细致地探索，寻找新的相移向量$\boldsymbol{\theta}'$。由于当前智能体训练尚不充分，其所认为的“最优臂”仍然存在进一步优化的可能性，因此，将臂空间内“最优臂”附近的区域称为“更有希望的区域”，通过在基准向量上加上细微的波动，实现对基准向量的微调，增强了对更有希望的臂空间区域的探索。可见，在用户位置几乎不变的通信场景下，该方法可以显著缩小探索范围，增加探索最优臂的概率。

3.2　线性估计补充法

在初始化阶段，离散出的臂集无法充分覆盖整个臂空间，因此，需要通过其他方法向臂集中添加新的向量作为新臂进行训练。基于LinUCB算法框架[25]，本文提出了线性补充估计法设计相移向量。该方法主要利用了信道能量与相移向量之间的线性关系，并基于历史的相移向量选择数据，估计出准确的信道协方差矩阵，据此设计出一个相移向量进行通信。

对于所有用户，训练能量是各用户k与基站间信道能量，即：

$$r_k(\boldsymbol{\theta}_t)=E\left(\hat{\boldsymbol{h}}_k(t)\hat{\boldsymbol{h}}_k^{\mathrm{H}}(t)\right) \tag{17}$$

$$r(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}E\left(\hat{\boldsymbol{h}}_k(t)\hat{\boldsymbol{h}}_k^{\mathrm{H}}(t)\right) \tag{18}$$

将有效信道的估计量式（11）代入训练能量式（18），则训练能量也可写成：

$$r(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}\left(E\left(\bar{\boldsymbol{h}}_k(t)\bar{\boldsymbol{h}}_k^{\mathrm{H}}(t)\right)+E\left(\boldsymbol{n}_k(t)\boldsymbol{n}_k^{\mathrm{H}}(t)\right)\right) \tag{19}$$

其中，用户k的级联信道为：

$$\bar{\boldsymbol{h}}_k(t)=\boldsymbol{h}_k^{\mathrm{H}}(t)\,\mathrm{diag}(\boldsymbol{\theta}_t)\boldsymbol{G}(t)=\boldsymbol{\theta}_t\mathrm{diag}\left(\boldsymbol{h}_k^{\mathrm{H}}(t)\right)\boldsymbol{G}(t) \tag{20}$$

令$\boldsymbol{H}_k(t)=\mathrm{diag}\left(\boldsymbol{h}_k^{\mathrm{H}}(t)\right)\boldsymbol{G}(t)$，训练能量表达形式变为：

$$r(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}E\left(\left(\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{H}_k(t)\right)\left(\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{H}_k(t)\right)^{\mathrm{H}}\right)+\sum_{k=1}^{K}E\left(\boldsymbol{n}_k(t)\boldsymbol{n}_k^{\mathrm{H}}(t)\boldsymbol{I}\right) \tag{21}$$

$$r(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}E\left(\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{H}_k(t)\boldsymbol{H}_k^{\mathrm{H}}(t)\boldsymbol{\theta}_t^{*}\right)+\sum_{k=1}^{K}E\left(\boldsymbol{n}_k(t)\boldsymbol{n}_k^{\mathrm{H}}(t)\boldsymbol{I}\right) \tag{22}$$

由于相移向量各元素模长为1，即$\left|\boldsymbol{\theta}_{i,t}\right|=1,i=1,2,\cdots,N$，对于$N$维的复数向量而言，其自相关表示为$\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\theta}_t^{*}=N$，则训练能量的表达式为：

$$r(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}\left(\boldsymbol{\theta}_t^{\mathrm{T}}E\left(\boldsymbol{H}_k(t)\boldsymbol{H}_k(t)^{\mathrm{H}}+1/N\,\boldsymbol{n}_k(t)\boldsymbol{n}_k^{\mathrm{H}}(t)\boldsymbol{I}\right)\boldsymbol{\theta}_t^{*}\right) \tag{23}$$

本文目标是设计相移向量$\boldsymbol{\theta}$，因此将信道协方差矩阵的部分记作一个整体$\boldsymbol{R}$，训练能量的表达式为：

$$r(\boldsymbol{\theta}_t)=\boldsymbol{\theta}_t^{\mathrm{T}}E\left(\sum_{k=1}^{K}\left(\boldsymbol{H}_k(t)\boldsymbol{H}_k(t)^{\mathrm{H}}\right)+1/N\sum_{k=1}^{K}\boldsymbol{n}_k(t)\boldsymbol{n}_k^{\mathrm{H}}(t)\boldsymbol{I}\right)\boldsymbol{\theta}_t^{*}=\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{R}\boldsymbol{\theta}_t^{*} \tag{24}$$

将训练能量$r(\boldsymbol{\theta}_t)$进行向量化后，可以得到新的线性形式[26]：

$$r'(\boldsymbol{\theta}_t)=\left(\boldsymbol{\theta}_t^{\mathrm{H}}\otimes\boldsymbol{\theta}_t^{\mathrm{T}}\right)\cdot\mathrm{vec}(\boldsymbol{R}) \tag{25}$$

其中，$\otimes$为克罗内克（Kronecker）积，vec(·)为向量化符号。基于这种线性形式，智能体可以在LinUCB算法框架下进行训练，以估计训练能量式（25）中的信道状态向量vec($\boldsymbol{R}$)。定义$\boldsymbol{\theta}_t^{\mathrm{H}}\otimes\boldsymbol{\theta}_t^{\mathrm{T}}$为相移向量$\boldsymbol{\theta}_t$的相关信息，也称为上下文信息$\boldsymbol{x}_t^{\mathrm{T}}$[27]。根据式（25），臂集内各臂的预期训练能量与上下文信息呈线性关系，其线性函数为：

$$E\left(r_t|\boldsymbol{x}_t\right)=\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\delta}^{*} \tag{26}$$

其中，$\boldsymbol{\delta}^{*}$为系数向量，是需要估计的未知参数，对应信道状态向量vec($\boldsymbol{R}$)。

在前t个时隙内，基站接收到t个从用户端发来的信号。由上下文向量组成的矩阵为$\boldsymbol{D}_t$，即$\boldsymbol{D}_t=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t\right]^{\mathrm{T}}$，维度是$t\times N^2$。在$t$个时间段内观察到用户$k$的实际能量为$\boldsymbol{c}_{t,k}$，即$\boldsymbol{c}_{t,k}=\left[r_{k,1},r_{k,2},\cdots,r_{k,t}\right]^{\mathrm{T}}$，维度是$t\times1$，其中，$r_{k,t}$是$t$时刻基站与用户$k$通信后的信道能量，即$r_{k,t}=\boldsymbol{y}_k(t)\boldsymbol{y}_k^{\mathrm{H}}(t)$。因此，根据接收到的能量$\boldsymbol{c}_{t,k}$和上下文向量矩阵$\boldsymbol{D}_t$，采用最小二乘法，通过求解如下的平方损失函数的最小值，能够得到未知向量系数$\boldsymbol{\delta}^{*}$的最优解：

$$\mathrm{loss}=\left(\boldsymbol{c}_t-\boldsymbol{D}_t\boldsymbol{\delta}^{*}\right)^2+\lambda\left\|\boldsymbol{\delta}^{*}\right\| \tag{27}$$

其中，λ为正则项系数。将loss函数对$\boldsymbol{\delta}^{*}$求导，即：$\nabla_{1\,\mathrm{oss}}=2\boldsymbol{D}_t^{\mathrm{T}}\left(\boldsymbol{c}_t-\boldsymbol{D}_t\boldsymbol{\delta}^{*}\right)-2\lambda\boldsymbol{\delta}^{*}$。令$\nabla_{\mathrm{loss}}=0$、$\lambda=$

1，则$\boldsymbol{\delta}^*$的表达式为式（28）～式（31）：

$$\boldsymbol{\delta}^*=\sum_{k=1}^{K}\boldsymbol{\delta}_k^*=\sum_{k=1}^{K}\boldsymbol{A}_t^{-1}\boldsymbol{b}_{t,k} \tag{28}$$

$$\boldsymbol{A}_t=\boldsymbol{D}_t^{\mathrm{T}}\boldsymbol{D}_t+\boldsymbol{I}_d \tag{29}$$

$$\boldsymbol{b}_t=\sum_{k=1}^{K}\boldsymbol{D}_t^{\mathrm{T}}\boldsymbol{c}_{t,k} \tag{30}$$

$$\boldsymbol{c}_t=\sum_{k=1}^{K}\boldsymbol{c}_{t,k} \tag{31}$$

其中，$\boldsymbol{I}_d$为单位矩阵，维度为$N^2\times N^2$。

在迭代过程中，上下文矩阵$\boldsymbol{A}_t$和训练数据矩阵$\boldsymbol{b}_t$的更新为：

$$\boldsymbol{A}_t=\boldsymbol{A}_{t-1}+\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}} \tag{32}$$

$$(\boldsymbol{A}_t)^{-1}=(\boldsymbol{A}_{t-1}+\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}})^{-1}=(\boldsymbol{A}_{t-1})^{-1}-\frac{(\boldsymbol{A}_{t-1})^{-1}\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}(\boldsymbol{A}_{t-1})^{-1}}{1+\boldsymbol{x}_t^{\mathrm{T}}(\boldsymbol{A}_{t-1})^{-1}\boldsymbol{x}_t} \tag{33}$$

$$\boldsymbol{b}_t=\boldsymbol{b}_{t-1}+\sum_{k=1}^{K}r_{k,t}\boldsymbol{x}_t \tag{34}$$

式（33）可由谢尔曼–莫里森（Sherman-Morrison）公式[28]获得。

各时隙内，智能体选择臂后，计算其上下文矩阵以及训练数据矩阵，并根据式（28）计算出系数向量$\boldsymbol{\delta}^*$，通过逆向量化后，便能得到估计的信道协方差矩阵$\hat{\boldsymbol{R}}=\mathrm{unvec}(\boldsymbol{\delta}^*)$，其中unvec(·)为逆向量化符号。随着时隙数量的增加，上下文信息以及训练数据不断丰富，因此，通过上下文信息及历史训练数据估计的信道协方差矩阵将会变得更加准确。

结合训练能量式（24），估计出信道协方差矩阵后，希望通过信道能量来设计RIS相移向量。由于最大化有效信道能量往往会使SINR随之增大，从而提高有效频谱效率。因此，当准确估计出了信道协方差矩阵时，可以通过如下目标求解次优臂：

$$\mathrm{P}(2)\quad \max_{\boldsymbol{\theta}}\quad \boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{\theta}^* \quad \text{s.t.}\ \ \left|[\boldsymbol{\theta}]_n\right|=1,\ \ n=1,\cdots,N \tag{35}$$

由于标量值的迹就是其本身，并根据迹的循环性质，有如下转换：$\boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{\theta}^*=\mathrm{tr}\left(\boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{\theta}^*\right)=tr(\hat{\boldsymbol{R}}\boldsymbol{\theta}^*\boldsymbol{\theta}^{\mathrm{T}})$，其中，$t(\cdot)$代表求矩阵的迹。因此，定义$\boldsymbol{V}=\boldsymbol{\theta}^*\boldsymbol{\theta}^{\mathrm{T}}$，其中$\boldsymbol{V}\succeq 0$且rank($\boldsymbol{V}$)=1，则P(2)命题等价于：

$$\mathrm{P}(3)\quad \max_{\boldsymbol{V}}\quad \mathrm{tr}(\hat{\boldsymbol{R}}\boldsymbol{V}) \quad \text{s.t.}\ \ \boldsymbol{V}\succeq 0,[\boldsymbol{V}]_{n,n}=1,\ \ n=1,\cdots,N,\ \ \mathrm{rank}(\boldsymbol{V})=1 \tag{36}$$

本文利用MATLAB的官方凸优化工具箱（convex optimization toolbox，CVX）来解决P(3)，在得到与相移向量相关的矩阵$\boldsymbol{V}$后，可以使用特征值分解（eigen value decomposition，EVD）法、高斯随机化[29]方法等来还原出相移向量$\boldsymbol{\theta}$。因此，线性补充估计法可以更新出一个基于有效信道能量最大化的RIS相移向量。

然而，利用该方法更新出的相移向量并不是最优臂。信号能量最大化并不能确保频谱效率最大化，这是因为有效频谱效率不仅与信号能量有关，还与干扰能量及噪声能量有关。此外，由于估计的信道不可避免地存在一定的误差，基于此设计的预编码与基于真实信道设计的预编码向量之间存在差异，实际的用户间干扰无法完全消除，无法稳定地获得能使有效频谱效率最大化的臂。然而，基于LinUCB算法框架的线性估计补充法设计的相移向量可以作为次优解加入臂集进行训练，以缩小在高维连续空间内的探索范围，从而加速整体算法的收敛过程。

在RIS辅助MIMO系统中，RIS反射单元数量以及基站天线数较大，这使得基站主动波束成形矩阵$\boldsymbol{W}$和RIS被动波束成形向量$\boldsymbol{\theta}$维度较大，因此，所提算法复杂度主要集中于线性补充估计法。层级贪婪搜索算法的复杂度上界为$O(K\times N^4)$，其中，通过式（28）、式（32）～式（34）更新信道协方差矩阵的复杂度为$O(K\times N^4)$，设计相移的复杂度为$O(N^{3.5})$，更新Q、N的复杂度均为$O(K)$，更新$\boldsymbol{A}$的复杂度为$O(N^4)$，更新$\boldsymbol{b}$的复杂度为$O(K\times N^2)$。在通信过程中，利用迫零

预编码设计基站主动波束成形的复杂度为$O(K^2\times M)$。由此可知，总的复杂度为$O(K\times N^4+K^2\times M)$。层级贪婪搜索算法如算法1所示。

算法1　层级贪婪搜索算法

输入　初始臂数 i，最大循环次数T

输出　RIS相移向量$\boldsymbol{\theta}$

在空间上随机生成臂集，个数为i；

循环

　初始化所有臂；

　根据式（15），更新均值奖励Q；

　根据式（16），更新选择次数N；

　根据式（29），更新包含上下文向量的矩阵$\boldsymbol{A}$；

　根据式（30），更新$\boldsymbol{b}$；

$t\leftarrow t+1$并且$t\leqslant i$；

循环

　随机取值ε，$\varepsilon\in(0,1)$；

　如果$\mathrm{e}^{-\mu t}\leqslant\varepsilon$，则：

　选择均值奖励Q最大的臂；

　　根据式（15），更新均值奖励Q；

　　根据式（16），更新选择次数N；

　　根据式（29），更新包含上下文向量的矩阵$\boldsymbol{A}$；

　　根据式（30），更新$\boldsymbol{b}$；

　如果$\mathrm{e}^{-\omega t}\leqslant\varepsilon<\mathrm{e}^{-\mu t}$，则：

　根据式（28）求解$\boldsymbol{R}$。通过$\boldsymbol{R}$设计一个相移向量$\boldsymbol{\theta}$；

　　根据式（15），更新均值奖励Q；

　　根据式（16），更新选择次数N；

　　根据式（29），更新包含上下文向量的矩阵$\boldsymbol{A}$；

　　根据式（30），更新$\boldsymbol{b}$；

　如果$\varepsilon<\mathrm{e}^{-\omega t}$，则：

　随机在均值奖励最大的附近臂选择一个相移向量$\boldsymbol{\theta}'$，即满足$\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\leqslant\alpha$；

　　根据式（15），更新均值奖励Q；

　　根据式（16），更新选择次数N；

　　根据式（29），更新包含上下文向量的矩阵$\boldsymbol{A}$；

　　根据式（30），更新$\boldsymbol{b}$；

当t达到最大循环次数T后，停止

4　仿真分析

本节通过MATLAB仿真，以验证所提方案的性能。本文的对比方法包括基于双时间尺度的级联有效信道增益最大化（maximization of cascaded effective channel gain，CECGM）[26,30]和随机选择RIS上相移向量的两种方法。CECGM方法的目标为级联有效信道增益最大化，通过交替方向乘子法估计的统计CSI信道协方差矩阵$\boldsymbol{R}$，并以最大化级联有效信道增益作为目标，优化相移向量。由于涉及的角度参数和路径增益的方差变化速度远低于瞬时CSI，因此，根据统计CSI设计出的RIS相移向量在一个长时间尺度内可以保持有效。随机选择方法则是随机选择相移向量。本文场景理论性能的界限在于，通过完美信道状态信息对RIS相移向量以及预编码进行设计，在完全消除用户间干扰的情况下，很好地利用RIS进行辅助通信。为了保持方案对比的公平性，上述方法均使用迫零预编码进行预编码设计。

4.1　参数设置

在每个时隙内的最大符号数设置为$S=100$，而导频符号数与基站天线数有关，随着基站天线数的增加而增加。基站采用$M_x\times M_y=8\times 8$天线的均匀平面阵列，RIS采用$N_x\times N_y=6\times 6$反射单元的平面阵列，用户数为$K=2$，且用户与基站和RIS的相对位置在大时间尺度内保持不变。基站的发射功率为$P_t=30$ dBm，在不同的信噪比（signal-to-noise ratio，SNR）下，噪声方差随SNR的变化进行调整。基站到RIS的信道和RIS到用户的信道分别根据式（2）和式（4）生成。

路径数设置为$L=P=3$，每个通道包括1条视距路径和2条非视距路径。信道为视距路径与非视距路径的复增益，分别满足$\alpha_1 \sim \mathcal{CN}(0,1)$，$\alpha_i \sim \mathcal{CN}(0,0.1), i=2,\cdots,P$，$\mu$、$\omega$分别设置为0.005和0.01，初始臂数$i=100$。

4.2 仿真结果分析

当基站天线数为36，RIS反射元件数为36时，有效频谱效率与SNR的关系如图3所示。从图3可以观察到，随着SNR的增加，3种方法的有效频谱效率都随之增加。整体上，本文提出的方法效果优于另外两种方案。此外，当SNR较低时，本文提出的方法有效频谱效率与对比方案性能相当。这是因为，在SNR较低的场景下，选择同一个臂时，其均值奖励因噪声的干扰而波动较大，这极大地降低了所提算法的收敛效率。在信道环境极差的情况下，噪声方差大，当选择次数不充足时，其均值奖励与实际奖励差距较大，因此，所提算法可能会表现不佳。然而，随着SNR的增加，当通信条件越来越好时，本文所提出的层级贪婪搜索算法的优势愈发明显。此外，由于噪声影响以及开始时隙内的探索，相较于基站已知完美CSI设计出的RIS被动波束成形矩阵以及预编码矩阵稍显不足。与Greedy、ε-Greedy策略相比，本文所提算法在相同初始臂集的情况下，通过线性补充估计法和随机收缩探索法，极大地提高了ε-Greedy策略的性能。

当基站天线数为36，RIS反射元件数为64时，有效频谱效率与SNR之间的关系如图4所示。与图3相比，仿真中将RIS反射元件数增至64，其余仿真参数保持不变。由于RIS数量的增加，在相同SNR的条件下，有效频谱效率得到了提升，并且本文所提出的层级贪婪搜索算法相较于另外两种方法以及传统Greedy、ε-Greedy策略仍然能够保持优势。其原因在于，本文所提算法对ε-Greedy策略进行了改进，降低了估计瞬时CSI所需的导频开销，并且在增加RIS数的情况下，通过层级贪婪搜索，为训练提供了更优质的RIS相移。此外，随着SNR的增加，噪声对主被动波束成形矩阵的影响越来越小，因此，所提层级贪婪搜索算法的性能也逐渐接近性能上限。

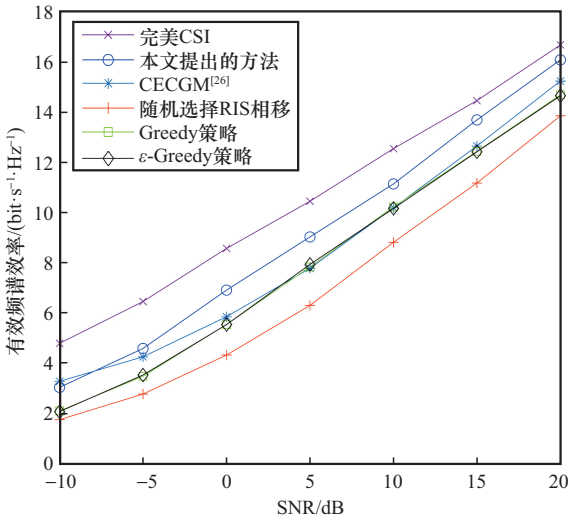


图3 有效频谱效率与SNR的关系

（基站天线数为36，RIS反射元件数为36）

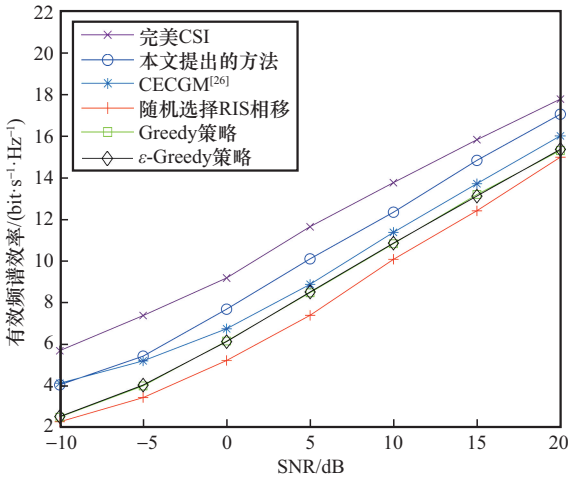


图4 有效频谱效率与SNR的关系

（基站天线数为36，RIS反射元件数为64）

当基站天线数为64，RIS反射元件数为36时，有效频谱效率与SNR的关系如图5所示。相较于图3中的场景，随着基站天线数的增加，导频开销也随之增加。在一个相干时间内，由于可用总符号数有限，有效传输时间随着天线数量的增加而减少。因此，在相同RIS数下，可能会因为导频开销的增

加，降低有效频谱效率。从图5可以观察到，与其余两种方案相比，所提出的层级贪婪搜索算法仍然有较大的优势，在相同的SNR条件下，层级贪婪搜索算法优于另外两种方案，并且随着SNR的增大优势也相对扩大。随着SNR的增加，所提算法的性能也逐渐趋于完美CSI设计的主被动波束成形性能。此外，虽然Greedy、ε-Greedy策略能够达到与传统方法相当的性能，但是，由于本文对ε-Greedy算法进行了改进，因此，Greedy、ε-Greedy策略始终落后于本文所提算法。

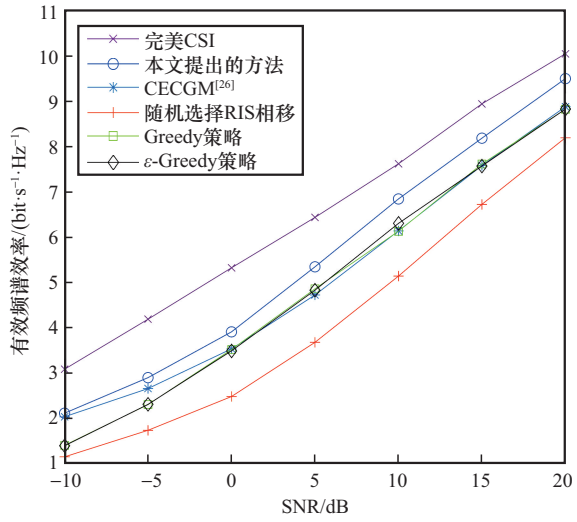


图5　有效频谱效率与SNR的关系
（基站天线数为64，RIS反射元件数为36）

发射功率与有效频谱效率的关系如图6所示。图6探讨了在相同场景下，不同发射功率对有效频谱效率的影响。在基站天线数为36，RIS上反射单元数为36，噪声服从复高斯分布$\mathcal{CN}(0,0.01)$的场景下，随着发射功率的提高，所得到的有效频谱效率也随之提高。本文所提算法与传统方法始终保持优势，并与理论上限保持着几乎恒定的差距。此外，Greedy、ε-Greedy策略虽然也有较低的开销，能够在一定程度上提高性能，但是由于直接离散方法生成的臂集容易错失最优臂，因此，ε-Greedy策略性能逊于本文所提算法。上述对比再次论证了所提方案的优越性。

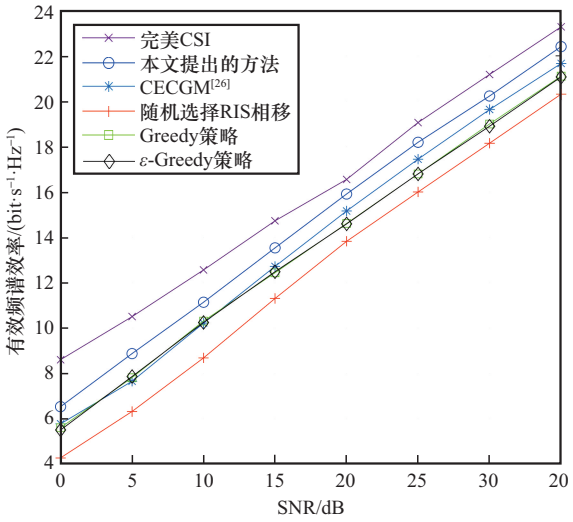
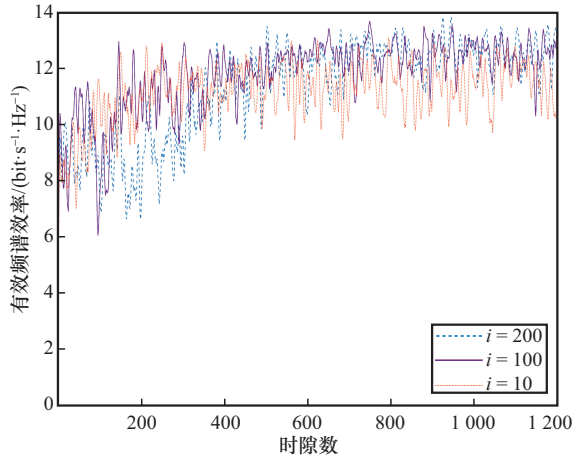


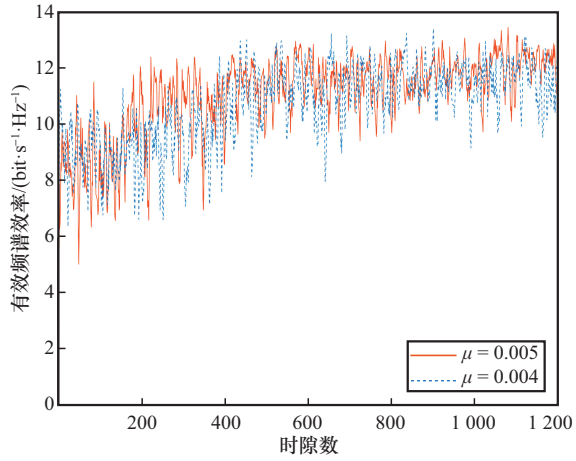
图6　发射功率与有效频谱效率的关系

本文所提算法在不同条件下的有效频谱效率随时隙变化的收敛图如图7所示。本文所提算法收敛速度与初始臂数有关，当初始臂数$i=10$时，收敛速度最快，在大约175个时隙后就已经收敛，但是由于初始化阶段收集的历史信道信息不充分，在利用次优补充估计法探索时未能探索到质量较优的臂，因此，收敛后的有效频谱效率相对较低。初始臂数增加至$i=100$时，大约240个时隙后收敛；初始臂数增加至$i=200$时，大约400个时隙后收敛。由于初始化占用部分时隙，因此，收敛速度随初始臂数的增加而减小。本文所提算法的收敛速度与探索因子有关，从图7（b）可以观察到，$\mu=0.004$时，大约在400个时隙后收敛，$\mu=0.005$时，大约在200个时隙后收敛，可见，改变探索概率因子会影响算法的收敛效率。当减小μ的值时，即在算法初期减小了利用概率，此时算法在初期会更多地进行探索，进而降低了算法的收敛效率，当增大探索概率因子时，则会在更多的时隙上进行利用。本文所提算法收敛速度与SNR有关，当SNR为20时，几乎在175个时隙后就已经探索到了最优臂，并进行利用。当SNR为0时，几乎在300个时隙后才收敛。由于不同时隙内，信道小尺度衰落不同，选择同一个

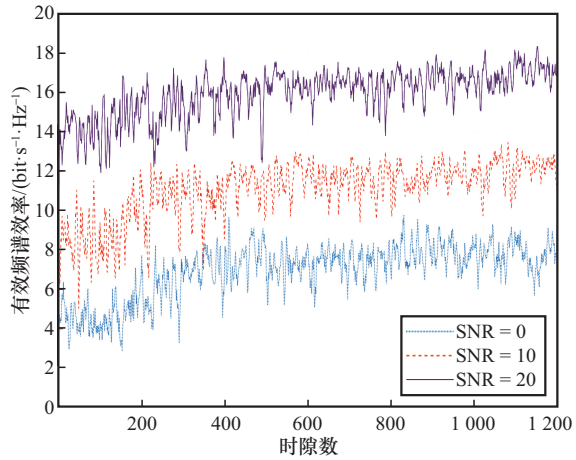
臂会使其获得的有效频谱效率波动较大，因此，其收敛速度较慢。总的来说，收敛速度与SNR、初始臂数以及探索概率因子有关。



（a）不同初始臂数



（b）不同探索概率



（c）不同SNR

图7　本文所提算法在不同条件下的有效频谱效率随时隙变化的收敛图

观察初始臂数$i=100$、$\mu=0.005$的收敛情况。在开始的一段时隙内，首先初始化臂集中所有的臂，该阶段平均有效频谱效率相对较低。本文利用历史时隙内选择的RIS相移向量构成的上下文信息进行进一步探索与设计，在MAB框架下，逐渐训练出更为准确的信道协方差矩阵。同时，在整个探索与利用阶段，通过最优贪婪法选择出最适合当前场景下的RIS相移向量。在大约200个时隙后，几乎已经探索到最优的相移向量。收敛后，有效频谱效率的值仍然会小幅波动，主要原因在于存在小尺度衰落和噪声的影响，即使RIS上的相移未发生改变，当前系统的有效频谱效率也会改变。此外，即使收敛后，也会进行一定的探索，防止陷入局部解，但是随着时隙数的增加，探索的概率会逐步减少。因此，后续时隙中虽然会有极少部分有效频谱效率较低，但也正因如此，其置信上限相对于臂集内的其他臂较小，再次选择的概率就变得更小。整体而言，层级贪婪搜索算法中的随机探索法对平均有效频谱效率的影响有限，但却增加了探索到更优秀的臂的概率。

5　结束语

本文考虑了RIS辅助下的多用户MIMO传输系统，提出了层级贪婪搜索算法，利用历史信息设计RIS相移向量，优化了基站处的预编码向量，最终达到了系统有效频谱效率的最大化，并将其进行仿真对比。仿真结果表明，无论是基站天线数改变、RIS反射元件数改变，还是发射功率发生变化，所提方案均展现出一定优势，这凸显了利用上下文信息的MAB方法在训练相移向量、减小导频开销方面的显著效果。在后续的工作中，将继续探寻RIS相移向量的设计方法，或将引入深度学习方法来进行训练。

参考文献：

[1]　Andrews J G, Buzzi S, Choi W, et al. What will 5G be?[J]. IEEE

Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[2] Al-Shuwaili A, Zaki N D, Abed G S, et al. Channel characterization for RIS-enabled indoor mmWave communications[C]//Proceedings of the 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA). Piscataway: IEEE Press, 2022: 89-93.

[3] Wang R Z, Ren H, Pan C H, et al. Channel estimation for RIS-aided mmWave massive MIMO system using few-bit ADCs[J]. IEEE Communications Letters, 2023, 27(3): 961-965.

[4] Dash S P, Kaushik A. RIS-assisted 6G wireless communications: a novel statistical framework in the presence of direct channel[J]. IEEE Communications Letters, 2024, 28(3): 717-721.

[5] 朱路虎, 王安定 . 一种基于ADMM的多用户联合的RIS信道估计方案[J]. 电信科学, 2024, 40(12): 74-85.
Zhu L H, Wang A D. Multi-user joint RIS channel estimation based on ADMM[J]. Telecommunications Science, 2024, 40(12): 74-85.

[6] Liu Y W, Liu X, Mu X D, et al. Reconfigurable intelligent surfaces: principles and opportunities[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1546-1577.

[7] Dai J X, Zhang S L, Zhi K D, et al. Two-timescale design for simultaneous transmitting and reflecting RIS-assisted massive MIMO systems with imperfect CSI[J]. IEEE Transactions on Communications, 2024, 72(7): 4287-4304.

[8] Basar E, Di Renzo M, De Rosny J, et al. Wireless communications through reconfigurable intelligent surfaces[J]. IEEE Access, 2019, 7: 116753-116773.

[9] Zhang Z J, Dai L L. Reconfigurable intelligent surfaces for 6G: nine fundamental issues and one critical problem[J]. Tsinghua Science and Technology, 2023, 28(5): 929-939.

[10] Chen J C. Joint transceiver and intelligent reflecting surface design for mmWave massive MIMO systems[J]. IEEE Systems Journal, 2023, 17(1): 792-803.

[11] Di Renzo M, Danufane F H, Tretyakov S. Communication models for reconfigurable intelligent surfaces: from surface electromagnetics to wireless networks optimization[J]. Proceedings of the IEEE, 2022, 110(9): 1164-1209.

[12] Wu Q Q, Zhang R. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5394-5409.

[13] Luo H H, Liu R, Li M, et al. Joint beamforming design for RIS-assisted integrated sensing and communication systems[J]. IEEE Transactions on Vehicular Technology, 2022, 71(12): 13393-13397.

[14] He Z Q, Yuan X J. Cascaded channel estimation for large intelligent metasurface assisted massive MIMO[J]. IEEE Wireless Communications Letters, 2020, 9(2): 210-214.

[15] Zhao M M, Wu Q Q, Zhao M J, et al. Intelligent reflecting surface enhanced wireless networks: two-timescale beamforming optimization[J]. IEEE Transactions on Wireless Communications, 2021, 20(1): 2-17.

[16] Cao Y S, Lv T J, Ni W. Two-timescale optimization for intelligent reflecting surface-assisted MIMO transmission in fast-changing channels[J]. IEEE Transactions on Wireless Communications, 2022, 21(12): 10424-10437.

[17] Zhi K D, Pan C H, Ren H, et al. Two-timescale design for reconfigurable intelligent surface-aided massive MIMO systems with imperfect CSI[J]. IEEE Transactions on Information Theory, 2023, 69(5): 3001-3033.

[18] Sutton R S. Reinforcement learning: an introduction[M]. Cambridge, Massachusetts: MIT Press, 1998.

[19] Qian M, Li C, Ma Y, et al. A Contextual MAB-based two-timescale scheme for RIS-assisted systems[J]. IEEE Wireless Communications Letters, 2025, 14(2): 400-404.

[20] Ayach O E, Rajagopal S, Abu-Surra S, et al. Spatially sparse precoding in millimeter wave MIMO systems[J]. IEEE Transactions on Wireless Communications, 2014, 13(3): 1499-1513.

[21] Minn H, Al-Dhahir N. Optimal training signals for MIMO OFDM channel estimation[J]. IEEE Transactions on Wireless Communications, 2006, 5(5): 1158-1168.

[22] Yang H, Marzetta T L. Performance of conjugate and zero-forcing beamforming in large-scale antenna systems[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(2): 172-179.

[23] Lattimore T, SzepesváRi C. Bandit algorithms[M]. Cambridge: Cambridge University Press, 2020: 75-83.

[24] El Jaghaoui S, Elmiad A K, Lmah A B. Enhancing the traveling salesman problem solutions with reinforcement learning: a variant exploration-exploitation approach beyond ε -Greedy[C]// Proceedings of 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA). Casablanca, Morocco: IEEE, 2023: 1-6.

[25] Li L, Chu W, Langford J, et al. A contextual-bandit approach to personalized news article recommendation[C]//Proceedings of the 19th International Conference on World Wide Web. Ra-

leigh, North Carolina, USA: ACM, 2010: 661-670.

[26] Wang H, Fang J, Duan H, et al. Spatial channel covariance estimation and two-timescale beamforming for IRS-assisted millimeter wave systems[J]. IEEE Transactions on Wireless Communications, 2023, 22(9): 6048-6060.

[27] Su Y, Xiong D, Wan Y, et al. LinFuzz: program-sensitive seed scheduling Greybox fuzzing based on LinUCB algorithm[J]. IEEE Access, 2024, 12: 74843-74860.

[28] Chen J, Zhao L, Jiang M, et al. Sherman-morrison formula aided adaptive channel estimation for underwater visible light communication with fractionally-sampled OFDM[J]. IEEE Transactions on Signal Processing, 2020, 68: 2784-2798.

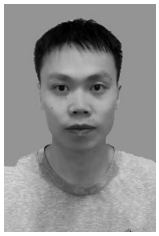
[29] So A M C, Zhang J, Ye Y. On approximating complex quadratic optimization problems via semidefinite programming relaxations[J]. Mathematical Programming, 2007, 110(1): 93-110.

[30] Wang P, Fang J, Wu Z, et al. Two-timescale beamforming for IRS-assisted millimeter wave systems: a deep unrolling-based stochastic optimization approach[C]//Proceedings of 2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM). Trondheim, Norway: IEEE, 2022: 191-195.

[作者简介]

沈天泽（2001– ），男，南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院硕士生，主要研究方向为多臂赌博机、智能反射面等。

汪革（1994– ），女，南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院博士生，主要研究方向为无人机通信、多臂赌博机等。



宋云超（1988– ），男，博士，南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院副教授、硕士生导师，主要研究方向为无线通信信号处理、机器学习等。

高天宝（1994– ），男，南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院博士生，主要研究方向为无线通信信号处理。

梁汇彬（1998– ），男，南京邮电大学电子与光学工程学院、柔性电子（未来技术）学院博士生，主要研究方向为大规模MIMO通信、多臂赌博机、卫星通信等。