Online Learning-Based Beamforming for Rate-Splitting Multiple Access: A Constrained Bandit Approach

Shangshang Wang, Jingye Wang, Yijie Mao, Ziyu Shao
School of Information Science and Technology, ShanghaiTech University
Email: {wangshsh2, wangjy5, maoyj, shaozy}@shanghaitech.edu.cn

***Abstract*—Rate-splitting multiple access (RSMA) has emerged as a potential non-orthogonal transmission strategy and powerful interference management scheme for 6G. Most of the existing works on RSMA beamforming design assume instantaneous or statistical channel state information (CSI) is available at the transmitter. Such assumption however is impractical especially in massive multiple-input multiple-output (MIMO) due to the dynamic wireless environments and the challenges in channel estimation. In this work, we propose a novel beamforming design framework based on online learning and online control to adaptively learn the best precoding action for a RSMA-aided downlink massive MIMO without explicit CSI feedback. In particular, we first formulate the precoder selection problem that maximizes the ergodic sum-rate subject to a long-term transmit power constraint as a constrained combinatorial multi-armed bandit (CMAB) problem. Then we propose a precoder selection with bandit learning algorithm for RSMA (PBR). Our theoretical analysis shows that PBR achieves a sublinear regret bound with a long-term power constraint guarantee. Through experimental results, we not only verify our theoretical analysis but also demonstrate the outperformance of PBR in terms of sum-rate and power consumption compared with the conventional transmission schemes without using RSMA.**

I. Introduction

Rate-splitting multiple access (RSMA) has shown its great potential in enhancing the spectral efficiency (SE), energy efficiency (EE), coverage, user fairness, reliability, and quality of service (QoS) for the next-generation communication systems [1]. Unlike the conventional linearly-precoded spatial division multiple access (SDMA) and power-domain non-orthogonal multiple access (PD-NOMA) which directly encode user messages into the corresponding data streams, in RSMA, each user message is split into a common part and a private part. The common parts of all users are encoded into a common stream to be decoded by all users. The private part of each user is independently encoded into a private stream to be decoded at the corresponding user only after each user removing the common stream based on successive interference cancellation (SIC). By such means, RSMA enables to partially decode the interference and partially treat the interference as noise, which generalizes the linearly precoded SDMA, PD-NOMA, and orthogonal multiple access (OMA) [2].

The performance of RSMA is significantly influenced by the design of beamforming. The existing works on RSMA beamforming design can be generally divided into two categories, namely, low-complexity beamforming design [3]–[5] and beamforming optimization [5]–[8]. However, all the above works assume certain amount channel state information at the transmitter (CSIT) is known in advance. Specifically, in the case of perfect CSIT, Mao *et al.* proposed a weighted minimum mean squared error (WMMSE) algorithm [6] and a successive convex approximation (SCA) algorithm [9] to respectively solve the weighted sum-rate (WSR) and EE maximization problem. Zhou *et al.* [4] addressed the SE-EE tradeoff of RSMA based on weighted matched beamforming (MBF) and zero-forcing beamforming (ZFBF). In [3], [5], [8], related problems with imperfect CSIT were studied. In particular, Joudeh and Clerckx [5] adopted WMMSE to design the precoders that maximize the sum-rate. Dai *et al.* [3] proposed a general framework for precoder design and power allocation via weighted MBF and regularized ZF (RZF). Mao and Clerckx [8] proposed a non-linearly precoded RSMA scheme that outperforms dirty paper coding (DPC) when CSIT is imperfect. In [10], Yin *et al.* further investigated the performance of RSMA with statistical CSIT. All the above works assume perfect/imperfect instantaneous or statistical CSIT is available. Only Hieu *et al.* [11] investigated the power allocation of RSMA in multiple-input single-output (MISO) without CSIT based on deep reinforcement learning (DRL). Such approach however cannot be extended to the beamforming design.

In massive multiple-input multiple-output (MIMO) networks, CSIT is often unavailable due to fast mobile entities and high pilot contamination in channel estimation [12]. For beamforming design in massive MIMO systems without CSIT in advance, online learning methods have been investigated in many works. Leveraging DRL and game theory, Liu *et al.* [13] proposed multiple DRL-based algorithms to jointly solve beamforming selection and power allocation in general multiple access networks. In [14], Kim *et al.* utilized bandit learning for codebook-based precoder selection in downlink frequency-division duplexing (FDD) systems. Based on deep learning methods, Xu *et al.* [15] proposed a beamforming scheme for FDD systems that maximizes sum-rate without the need of channel state information (CSI) feedback. However, unlike the sum-rate maximization problem for SDMA, the problem for RSMA is typically more challenging to solve due to the non-smooth function for the rate of the common stream. Therefore, all the aforementioned works are not applicable to RSMA.

Motivated by the limitation of the existing works on RSMA and massive MIMO, we propose a novel beamforming design

framework based on online learning and online control to adaptively learn the best precoding action for a RSMA-aided downlink massive MIMO with only effective channel gain feedback. Our main contributions and key results are as follows.

- **Problem Formulation**: We formulate the online precoder selection problem of RSMA as a combinatorial multi-armed bandit (CMAB) problem, which maximizes the ergodic sum-rate while ensuring a long-term transmit power constraint.
- **Algorithm Design**: We propose a novel precoder selection with bandit learning algorithm for RSMA called PBR to solve the formulated problem. In PBR, we leverage bandit learning to capture the uncertainty of CSIT as well as Lyapunov optimization techniques to conduct online transmit power control.
- **Theoretical Analysis**: We provide rigorous theoretical analysis to show that PBR not only satisfies the long-term power constraint but also achieves a sublinear time-averaged regret bound.
- **Numerical Evaluation**: We conduct extensive experiments to evaluate the performance of PBR and its variants. The results demonstrate the effectiveness of PBR in achieving a higher sum-rate than the conventional SDMA-aided scheme.

The rest of this paper is organized as follows. In Section II, we introduce the system model and problem formulation. Section III elaborates the details of our algorithm design, followed by the theoretical analysis in Section IV. In Section V, we evaluate the performance of our algorithm via extensive numerical experiments. Finally, we conclude this paper in Section VI. All the proofs are delegated to our technical report [16].

Notation: Sets are denoted by calligraphic uppercase letters. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. $\mathbf{a}^H$ denotes the Hermitian transpose of vector $\mathbf{a}$. The operations $|\cdot|$ and $||\cdot||$ denote the absolute value of a scalar and l_2-norm of a vector, respectively. $\mathbf{I}$ denotes the identity matrix.

II. System Model and Problem Formulation

In this section, we illustrate the proposed transmission framework followed by the problem formulation for beamforming design.

A. Basic Model

We consider an FDD downlink massive MIMO system. The system operates on a finite time horizon of T time slots denoted by $\{0, 1, \ldots, T-1\}$ and consists of one base station (BS) equipped with L antennas and K single-antenna users indexed by set $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$.

Message: In each time slot t, the BS sends K messages for the K users, respectively. We denote the messages as $\mathbf{W}(t) = \{W_1(t), ..., W_K(t)\}$. With 1-layer RSMA [1], each message $W_k(t)$ is split into a common part and a private part, *i.e.,* $W_k(t) \to \{W_{\mathrm{c},k}(t), W_{\mathrm{p},k}(t)\}$. All common parts are encoded into a common stream to be decoded by all users, *i.e.,* $\{W_{\mathrm{c},1}(t), ..., W_{\mathrm{c},K}(t)\} \to s_{\mathrm{c}}(t)$, and the private parts

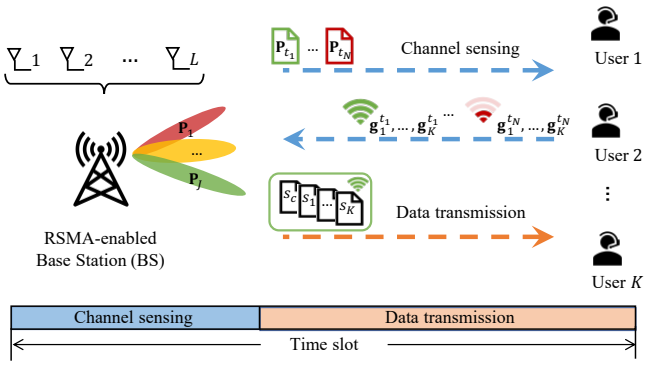

Fig. 1. The proposed RSMA-aided transmission framework for online beamforming design.

are encoded into private streams for the corresponding users independently, *i.e.,* $W_{\mathrm{p},k}(t) \to s_k(t)$. The stream vector to be transmitted is given by $\mathbf{s}(t) = [s_{\mathrm{c}}(t), s_1(t), ..., s_K(t)]^T \in \mathbb{C}^{(K+1)\times 1}$, where $\mathbb{E}\{\mathbf{s}(t)\mathbf{s}(t)^H = \mathbf{I}\}$.

Precoding: Let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_J\}$ denote the predefined codebook at the BS and $\mathbf{P}_i = [\mathbf{p}_{i,\mathrm{c}}, \mathbf{p}_{i,1}, ..., \mathbf{p}_{i,K}] \in \mathbb{C}^{L\times(K+1)}, i \in \mathcal{J} \triangleq \{1, 2, \ldots, J\}$, is composed of one set of candidate precoders. We use $\mathbf{P}(t) = [\mathbf{p}_{\mathrm{c}}(t), \mathbf{p}_1(t), ..., \mathbf{p}_K(t)]$ to denote a general precoding matrix selected in time slot t. Suppose the precoding matrix for encoding stream $\mathbf{s}(t)$ is $\mathbf{P}(t) \in \mathcal{P}$. Then the transmit signal $\mathbf{x}(t) \in \mathbb{C}^{L\times 1}$ is given by

$$\mathbf{x}(t) = \mathbf{P}(t)\mathbf{s}(t) = \mathbf{p}_{\mathrm{c}}(t)s_{\mathrm{c}}(t) + \sum_{i=1}^{K} \mathbf{p}_i(t)s_i(t). \tag{1}$$

Each time slot of the system is composed of two phases, namely, a channel sensing phase and a data transmission phase. In the channel sensing phase, the BS selects N precoding matrices from a predefined codebook, then sends each selected precoding matrix to all users to get the corresponding effective channel gains. In the data transmission phase, the BS serves K users based on RSMA with the optimal precoding matrix found in the channel sensing phase. We give a detailed description of the two phases in the following subsections. An illustration of the transmission framework is shown in Figure 1.

B. The Channel Sensing Phase

Let $\mathbf{h}_k(t) \in \mathbb{C}^{L\times 1}$ be the channel vector between the BS and user k at time slot t. Such a channel vector is assumed to be identically and independently distributed (*i.i.d.*) across time slots, and is usually unavailable at the BS [12]. Without loss of generality, we assume all users have the same additive white Gaussian noise (AWGN) $n_k(t) \sim \mathcal{CN}(0, 1)$. Then for each user k, the received signal $y_k(t) \in \mathbb{C}$ is given by

$$\begin{aligned} y_k(t) &= \mathbf{h}_k^H(t)\mathbf{x}(t) + n_k \\ &= \mathbf{h}_k^H(t)\mathbf{p}_{\mathrm{c}}(t)s_{\mathrm{c}}(t) + \sum_{i=1}^{K} \mathbf{h}_k^H(t)\mathbf{p}_i(t)s_i(t) + n_k \end{aligned} \tag{2}$$

In the channel sensing phase of time slot t, the BS needs to select N precoding matrices denoted by set $\mathcal{N}(t)$ from a predefined codebook, then sends each selected precoding matrix

to all users as illustrated in Figure 1. We denote the precoder selection decision by vector $\boldsymbol{I}(t) = (I_i(t))_{i\in\mathcal{J}}$ where $I_i(t) \in \{0,1\}$. If $I_i(t) = 1$, then precoding matrix $\mathbf{P}_i$ is selected in time slot t; $I_i(t) = 0$, otherwise. The precoder selection decision should satisfy the following constraints:

$$\begin{aligned} &\sum_{i\in\mathcal{J}} I_i(t) = N,\ \forall t, \\ &I_i(t) \in \{0,1\},\ \forall i \in \mathcal{J},\ \forall t. \end{aligned} \tag{3}$$

After receiving precoding matrix $\mathbf{P}_i$, user k observes the corresponding effective channel gain $\mathbf{g}_k^i(t) \triangleq [|\mathbf{h}_k^H(t)\mathbf{p}_{i,\mathrm{c}}|^2, |\mathbf{h}_k^H(t)\mathbf{p}_{i,1}|^2, ..., |\mathbf{h}_k^H(t)\mathbf{p}_{i,K}|^2] \in \mathbb{R}^{(K+1)}$ and sends it back to the BS. Based on the feedback, the BS computes the signal-to-interference-plus-noise ratio (SINR) and the rate for each user.

Common Rate: Given a selected precoding matrix $\mathbf{P}(t)$ and the corresponding effective channel gain, the SINR for the common stream of user k is

$$\gamma_{\mathrm{c},k}(\mathbf{P}(t)) = \frac{|\mathbf{h}_k^H(t)\mathbf{p}_{\mathrm{c}}(t)|^2}{\sum_{i=1}^K |\mathbf{h}_k^H(t)\mathbf{p}_i(t)|^2 + 1}, \tag{4}$$

and the common rate is given by

$$R_{\mathrm{c},k}(\mathbf{P}(t)) = \log_2(1 + \gamma_{\mathrm{c},k}(\mathbf{P}(t))). \tag{5}$$

To ensure all users are capable of decoding $s_{\mathrm{c}}(t)$, the common rate $R_{\mathrm{c}}(t)$ should not exceed the minimal common rate over all users, which yields

$$R_{\mathrm{c}}(\mathbf{P}(t)) \triangleq \min_{k\in\mathcal{K}} R_{\mathrm{c},k}(\mathbf{P}(t)). \tag{6}$$

Private Rate: Similarly, the SINR for private stream $s_k(t)$ of user k is

$$\gamma_{\mathrm{p},k}(\mathbf{P}(t)) = \frac{|\mathbf{h}_k^H(t)\mathbf{p}_k(t)|^2}{\sum_{i=1,i\neq k}^K |\mathbf{h}_k^H(t)\mathbf{p}_i(t)|^2 + 1}, \tag{7}$$

and the private rate is given by

$$R_{\mathrm{p},k}(\mathbf{P}(t)) = \log_2\left(1 + \gamma_{\mathrm{p},k}(\mathbf{P}(t))\right). \tag{8}$$

C. The Data Transmission Phase

In the data transmission phase of time slot t, the BS broadcasts k messages with the optimal precoding matrix in $\mathcal{N}(t)$ that achieves the maximal sum-rate.

Sum-Rate: For one selected precoding matrix $\mathbf{P}_i$, the sum-rate over both the common rate (6) and the private rate (8) is given as

$$R_i(t) \triangleq R_{\mathrm{c}}(\mathbf{P}_i) + \sum_{k=1}^K R_{\mathrm{p},k}(\mathbf{P}_i). \tag{9}$$

Based on the precoder selection decision $\boldsymbol{I}(t)$, the maximal sum-rate and the corresponding index of the optimal precoding matrix in time slot t are given by

$$R(t) \triangleq \max_{i\in\mathcal{J}} R_i(t)I_i(t), \tag{10}$$

$$\tilde{i}(t) \triangleq \operatorname*{argmax}_{i\in\mathcal{J}}\ R_i(t)I_i(t). \tag{11}$$

We further define $\tilde{\boldsymbol{I}}(t) \triangleq (\tilde{I}_i(t))_{i\in\mathcal{J}}$ to indicate the optimal precoding matrix selected by the BS in time slot t. In particular, for each precoding matrix $\mathbf{P}_i$, $\tilde{I}_i(t) = 1$ if $i = \tilde{i}(t)$ and 0 otherwise.

Remark: Note that in each time slot t, given a selected precoding matrix $\mathbf{P}_i$, the transmission rate $R_{\mathrm{c},k}(\mathbf{P}_i)$ and $R_{\mathrm{p},k}(\mathbf{P}_i)$ for user k are random variables due to the uncertainty of channel state information $\mathbf{h}_k(t)$. As $\mathbf{h}_k(t)$ is assumed to be *i.i.d.* across time slots, both $R_{\mathrm{c},k}(\mathbf{P}_i)$ and $R_{\mathrm{p},k}(\mathbf{P}_i)$ are *i.i.d.* across time slots, which yields *i.i.d.* $R_i(t)$ across time slots. We further denote the bound and the mean of $R_i(t)$ as $r_{\min} \leq R_i(t) \leq r_{\max}$ and μ_i. Thus the maximal sum-rate $R(t)$ is bounded and *i.i.d.* across time slots.

Transmit Power: For each precoding matrix $\mathbf{P}_i = [\mathbf{p}_{i,\mathrm{c}}, \mathbf{p}_{i,1}, ..., \mathbf{p}_{i,K}] \in \mathcal{P}$, the transmit power it would consumed at the BS is

$$P_i \triangleq \|\mathbf{p}_{i,\mathrm{c}}\|^2 + \sum_{k=1}^K \|\mathbf{p}_{i,k}\|^2, \tag{12}$$

which is assumed to be $0 \leq P_i \leq p_{\max}$. Then in time slot t, as the BS transmits data with the precoding matrix that achieves $R(t)$, the transmit power is given by

$$P(t) \triangleq \sum_{i\in\mathcal{J}} P_i\tilde{I}_i(t). \tag{13}$$

We consider the control of the transmit power from a long-term perspective for the system [17]. In particular, we have

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[P(t)\right] \leq p, \tag{14}$$

where p is a given constant power constraint over all time slots, and the constraint is ergodic over the selection decision $\boldsymbol{I}(t)$ made with unknown CSIT.

D. Problem Formulation

Our goal is to maximize the ergodic sum-rate under the long-term transmit power constraint at the BS over T time slots, which is defined by the following problem formulation:

$$\begin{aligned} \underset{\{\boldsymbol{I}(t)\}_t}{\text{maximize}}\ & \sum_{t=0}^{T-1} \mathbb{E}\left[R(t)\right] \\ \text{subject to}\ & (3)(14). \end{aligned} \tag{15}$$

III. Algorithm Design

As CSIT is usually unavailable at the BS in a massive MIMO system, we reformulate problem (15) from the perspective of constrained combinatorial multi-armed bandit (CMAB) [18]. Then we adopt Lyapunov optimization techniques [17] to decouple the reformulated problem into a series of sub-problems over time slots and propose an effective online scheme to solve each sub-problem. Below we first present our reformulation. Then we provide the details of our algorithm design.

A. Problem Reformulation

The classical CMAB [19] considers a sequential game between an agent and a bandit with multiple arms. The game is played over a finite number of time slots. In each time slot, the agent selects a subset of arms to play, then the agent receives

a reward with respect to each played arm. The objective of the agent is to maximize the expected cumulative reward over time slots.

For problem (16), the BS is viewed as the agent and the precoder set corresponds to the arm set. In each time slot t, the BS (agent) selects N precoding matrices from the set. After sending each selected precoding matrix to all users, the BS will receive the feedback about the corresponding effective channel gain. Based on the feedback, the BS computes the sum-rate according to (9). The maximal sum-rate is then viewed as reward $\mathrm{Rew}(t) \triangleq R(t)$. Therefore problem (15) can be rewritten as

$$\begin{aligned} \underset{\{\boldsymbol{I}(t)\}_t}{\text{maximize}} \quad & \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{Rew}\left(t\right)\right] \\ \text{subject to} \quad & (3)(14). \end{aligned} \tag{16}$$

We further define the regret $\mathrm{Reg}(T)$ that indicates the performance loss of the decision making under uncertainty as

$$\mathrm{Reg}\left(T\right) \triangleq \mathrm{Rew}^*(T) - \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{Rew}\left(t\right)\right], \tag{17}$$

where $\mathrm{Rew}^*(T)$ is the optimal expected reward that can be obtained given full knowledge of CSIT. Since $\mathrm{Rew}^*(T)$ is a constant, the reward maximization problem (16) is then can be reformulated as the following equivalent problem:

$$\begin{aligned} \underset{\{\boldsymbol{I}(t)\}_t}{\text{minimize}} \quad & \mathrm{Reg}\left(T\right) \\ \text{subject to} \quad & (3)(14). \end{aligned} \tag{18}$$

To solve problem (18), we integrate bandit learning [20] and Lyapunov optimization [17] to propose an online scheme called PBR (Precoder Selection with Bandit Learning for Rate-Splitting). We show the detailed algorithm design in the following subsections.

B. Online Learning for Uncertainty

We adopt the idea of online learning algorithm UCB [20] to deal with the uncertainty of CSIT. Let $z_i(t)$ be the number of times precoding matrix $\mathbf{P}_i$ is selected by time slot t, which is

$$z_i(t) = \sum_{\tau=0}^{t-1} I_i(\tau), \tag{19}$$

When the BS receives the effective channel gain of $\mathbf{P}_i$ at time slot t, the BS acquires new knowledge about its sum-rate $R_i(t)$. Denoting the empirical mean sum-rate $\bar{R}_i(t)$ of $\mathbf{P}_i$ over t time slots by

$$\bar{R}_i(t) = \frac{\sum_{\tau=0}^{t-1} R_i(\tau) I_i(\tau)}{z_i(t)}, \tag{20}$$

we have the estimation for its sum-rate as

$$\hat{R}_i(t) = \min\left\{\bar{R}_i(t) + \lambda\sqrt{\frac{\log t}{z_i(t)}}, r_{\max}\right\} \tag{21}$$

where λ is a positive parameter to balance the exploitation-exploration tradeoff. Specifically, the term $\bar{R}_i(t)$ is determined by the previous computation results about the sum-rate. The higher the value of $\bar{R}_i(t)$, the higher the estimation for the sum-rate of $\mathbf{P}_i(t)$, thus it can be viewed as the exploitation term. The second term $\sqrt{\log t / z_i(t)}$ is determined by the number of times precoding matrix $\mathbf{P}_i$ is selected. The less it has been selected, the higher the estimation, which gives it more chances to be selected in the next time slots. Therefore it can be regarded as the exploration term.

C. Online Control for Power Constraint

We define a virtual queue $Q(t)$ whose dynamics is given by

$$Q(t+1) \triangleq [Q(t) - p + P(t)]^+, \tag{22}$$

where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$ and $Q(0) = 0$. When the stability of queue $Q(t)$ is ensured, the power constraint (14) will be guaranteed [17]. To ensure the queue stability, we define

$$w_i(t) \triangleq Q(t) P_i, \tag{23}$$

which reflects the power consumption of precoding matrix $\mathbf{P}_i$ at the BS. The stochastic optimization problem (18) can then be transformed into a series of sub-problems over time slots [17], which is defined by

$$\begin{aligned} \underset{\boldsymbol{I}(t)}{\text{maximize}} \quad & V \max_{i:I_i(t)=1} \hat{R}_i(t) - \max_{i:I_i(t)=1} w_i(t) \\ \text{subject to} \quad & (3), \end{aligned} \tag{24}$$

where V determines the relative importance of sum-rate to transmit power. Intuitively, the objective of problem (24) is maximizing the sum-rate while minimizing the transmit power. Given a small value of V, the preference for precoders that entail a high transmit power is weak as the term $\sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t)$ is dominant. When the value of V is large, the preference for precoders with a high estimation of sum-rate is strong even though it may incur a high transmit power. The choice of V depends on the design of the system[1].

D. Online Scheme for Precoder Selection

Based on the above online learning and online control, we propose an online scheme for precoder selection called PBR. The pseudocode of PBR is given in Algorithm 1.

In each time slot, there are three procedures in PBR. The BS firstly estimates the sum-rate and computes the weight of each precoding matrix (lines 2-5). Then it selects N precoding matrices with the largest weights, and sends the selected precoding matrices to all users to get the corresponding effective channel gains (lines 6-15). After receiving the feedback, the BS finds the optimal precoding matrix and broadcasts messages with it, then updates the virtual queue (lines 16-18).

IV. Performance Analysis

In this section, we provide two theorems to analyze the performance of PBR.

[1]The objective of problem (24) is equivalent to maximizing $\max_{i:I_i(t)=1}(V\hat{R}_i(t) - w_i(t))$ when V is large enough, and the computation complexity of PBR can be greatly reduced in that case.

Algorithm 1 **P**recoder Selection with **B**andit Learning for **R**SMA (PBR)

Input: The parameter N which defines the number of precoding matrices selected in each time slot.
Output: Precoder selection decisions over T time slots, *i.e.,*$\{\boldsymbol{I}(t)\}_{t=0}^{T-1}$.
Initialization: The BS sends each precoding matrix $\mathbf{P}_i \in \mathcal{P}, i = 1, 2, ..., J$, to all users, and computes the corresponding sum-rate $R_i'(0)$ given the feedback of the effective channel gain, then sets $\bar{R}_i(0) \leftarrow R_i'(0)$, $z_i(0) \leftarrow 1$, and $Q(0) \leftarrow 0$.

for $t = 1, 2, \ldots, T-1$ **do**:
  % Weight Computation
  for each precoding matrix $\mathbf{P_i}, i = 1, 2, ..., J$ **do**
    Compute $\hat{R}_i(t)$ according to (21).
    Compute $w_i(t)$ according to (23).
  end for
  % Channel Sensing
  Set $R_i(t) = 0, i \in \mathcal{J}$.
  Get $\boldsymbol{I}(t)$ according to (24).
  for $i = 0, 1, ..., J$ **do**:
    if $I_i(t) = 1$ **then**
      Get the effective channel gain of precoding matrix $\mathbf{P}_i$.
      Compute $R_i(t)$ according to (9).
      Update $z_i(t)$ according to (19).
      Update $\bar{R}_i(t)$ according to (20).
    end if
  end for
  % Data Transmission
  Get $\tilde{i}(t)$ according to (11).
  Broadcast all messages $\mathbf{s}(t)$ with precoding matrix $\mathbf{P}_{\tilde{i}(t)}$.
  Update $Q(t)$ according to (22).
end for

A. Power Constraint Guarantee

A power constraint p is said to be feasible if there exists a feasible solution to problem (24) given the constraint p. The set of all feasible constraints is defined as the maximal feasibility region of problem (24).

Theorem 1: Suppose that the power constraint p lies in the interior of the maximal feasibility region, then PBR satisfies the power constraint in (14). Moreover, the virtual queue $Q(t)$ defined in (22) is strongly stable and there exists a positive constant ϵ such that

$$\limsup_{T' \to \infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}\left[Q(t)\right] \leq \frac{D + V r_{\max}}{\epsilon} < \infty,$$

where $D \triangleq \frac{1}{2} \max\left\{(p_{\max} - p)^2, p^2\right\}$ and V is a positive tunable parameter of finite value.

Remark 1: Theorem 1 shows that PBR can ensure the stability of the virtual queue, thereby can satisfy the time-averaged power constraint defined in (14) [17].

B. Regret Bound

Theorem 2: We have the following upper bound for the regret of PBR:

$$\mathrm{Reg}(T) \leq \frac{D}{V} + \frac{\Delta r + (\Delta r + 1) N H_\lambda}{T} + 4\lambda \sqrt{\frac{N \log T}{T}},$$

where $H_\lambda \triangleq \sum_{t=1}^{\infty} t^{-2\lambda^2}$ and $\Delta r \triangleq r_{\max} - r_{\min}$.

Remark 2: Theorem 2 shows that PBR achieves a regret bound of order $O(1/V + \sqrt{\log T / T})$, which is sub-linear given suitable values of V. For example, when the value of V is in order $\Omega(\sqrt{T}/\log T)$, the regret bound is sub-linear with order $O(\log T/\sqrt{T})$. Given a larger value of V (in the required order), PBR prefers minimizing the regret to ensuring the long-term constraint, hence the less the regret.

V. Numerical Results

In this section, we provide numerical results with corresponding discussion to verify the effectiveness of PBR.

A. Parameter Settings

We consider the number of antennas of the BS and the number of users to be $L = 64$ and $K = 4$, respectively. The long-term power constraint at the BS is set as $p = 20$ (W). In each time slot t, each element of the channel state information $\mathbf{h}_k[t]$ is sampled from complex Gaussian distribution $\mathcal{CN}(0, 1)$. The number of time slots is set as $T = 500$.

The predefined codebook $\mathcal{P}$ consists of 1000 precoding matrices, and the number of selected precoding matrices in each time slot is $N = 3$. For each precoding matrix $\mathbf{P}_i \in \mathbb{C}^{L \times (K+1)}$, we generate $L \times (K+1)$ complex numbers from $\mathcal{CN}(0, 1)$, then we normalize it so that $P_i = 1$ for further power allocation. The transmit power of each precoding matrix is sampled from uniform distribution $\mathcal{U}(17, 28)$ so that the BS can potentially achieve a higher sum-rate by selecting precoding matrices with higher transmit power. Note that in this paper, a long-term transmit power constraint p is considered, and therefore, the instantaneous transmit power consumption in each time slot could be higher than the threshold p. For the power allocation, we have 100 precoding matrices where all the corresponding transmit power is assigned to the common stream, and 100 precoding matrices where the corresponding transmit power is assigned to the private streams only. As for the remaining precoding matrices, the transmit power is randomly allocated for both the common stream and private streams.[2]

All the experiment results are averaged over 30 different random seeds. Unless otherwise stated, the value of V in the following experiments is $V = 20$, and when presenting power consumption related results, the time slots are ranging from 5 to 100 for better illustrations.

B. Performance Evaluation

1) Power Consumption of PBR over Time Slots: In Figure 2, we show how the power consumption from 5-th time slots evolves under different values of V. We see that the larger the value of V, the higher the power the BS consumes. However, in the long run, the time-averaged transmit power is controlled below the given constraint $p = 20$ (W), which is consistent with our theoretical analysis.

Besides, we also test PBR with different power constraints to further demonstrate its effectiveness in online control. The performance of PBR in terms of power consumption and sum-rate with power constraint $p \in \{18, 20, 22\}$ are shown in

[2]Note that our scheme is also applicable to other kinds of codebooks.

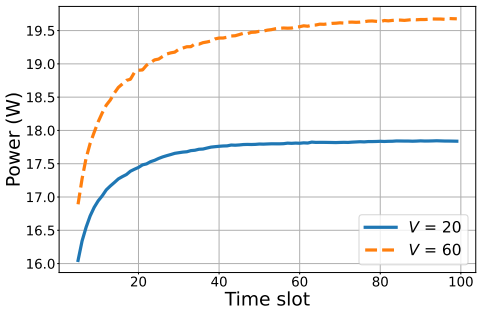

Fig. 2. Power consumption at the BS.

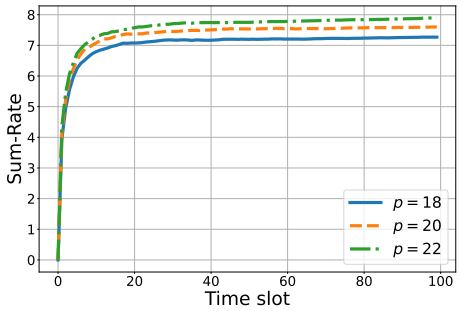

Fig. 3. Sum-rate of PBR with different power constraints p.

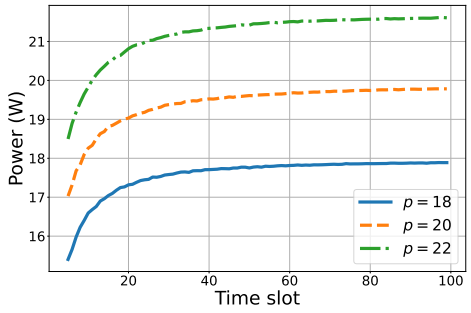

Fig. 4. Power consumption of PBR with different power constraints p.

Figure 3 and Figure 4, where the value of V is 60. The results indicate that on the one hand, given a higher power budget, PBR can achieve a higher sum-rate while ensuring the power constraint guarantee in the long term; on the other hand, when the power budget is low, PBR can also satisfy the constraint effectively at the cost of sum-rate. Specifically, when the power budget decreases by 10% (from 20 to 18), PBR can still ensure the power constraint guarantee with a loss of only 4.13% (from 7.8017 to 7.4796) in sum-rate.

2) Tradeoff between Regret and Power Consumption: In Figure 5, we show the tradeoff between regret and power consumption of PBR. Specifically, the larger the value of V, the lower the regret and the larger the power consumption. As suggested in our algorithm design, the tradeoff can be tuned by adjusting the value of V.

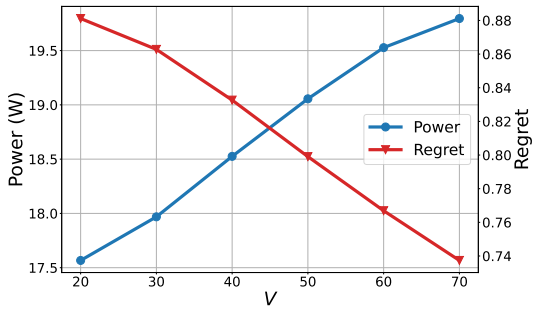

Fig. 5. Tradeoff under PBR.

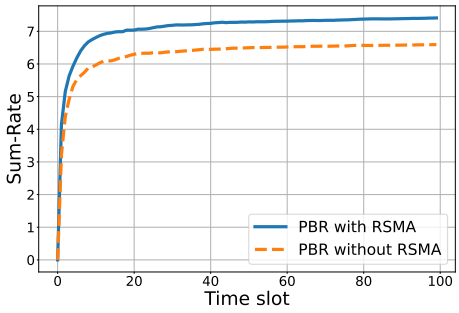

Fig. 6. Sum-rate of PBR with and without RSMA.

3) Sum-Rate of PBR with Rate-Splitting: In Figure 6, we show the sum-rate of PBR with and without RSMA. When RSMA is not used, the sum-rate in (9) will be composed of private rates only. The result shows that PBR can achieve a higher sum-rate when RSMA is taken into consideration. That is because in the RSMA case it can benefit from the power allocated for common rate. The curve of the sum-rate with RSMA is also consistent with our theoretical analysis. Besides, in Figure 7, we show the sum-rate of PBR with PBR and without PBR in the case of different number of users when $V = 40$. Both the sum-rates of the two cases increase as the number of users increase. However, the performance of PBR with RSMA is generally better than that of PBR without RSMA. The superiority of RSMA can also be demonstrated by the result shown in Figure 8 where $V = 40$. PBR can benefit from a higher power constraint to achieve a higher sum-rate, while PBR without RSMA has only little improvement.

In Figure 9, we further investigate the performance of PBR with RSMA in different scenarios when $V = 40$. Particularly, we test PBR with the cases where the numbers of users are $K = 6$ and $K = 8$. Compared with the sum-rate of $K = 4$, the sum-rates of $K = 6$ and $K = 8$ increase by 13.79% and 20.31% (from 7.7359 to 8.8029 and 9.3075, respectively), which demonstrates the effectiveness of RSMA and PBR in online learning.

4) Sum-Rate with Different Number of Selection: Recall that N is the number of precoding matrices selected in each time slot for channel sensing. In Figure 10, we show the sum-rate achieved by PBR with $N \in \{1, 3, 5\}$. The result indicates that

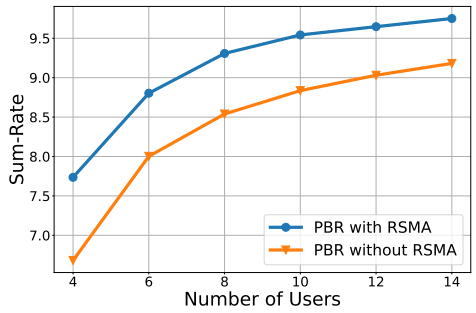

Fig. 7. Sum-rate of PBR with various number of users.

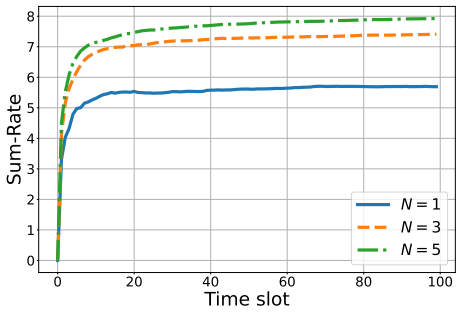

Fig. 10. Sum-rate of PBR under different values of N.

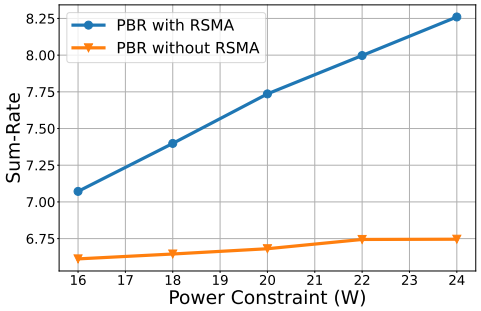

Fig. 8. Sum-rate of PBR with and without RSMA under different power constraints p.

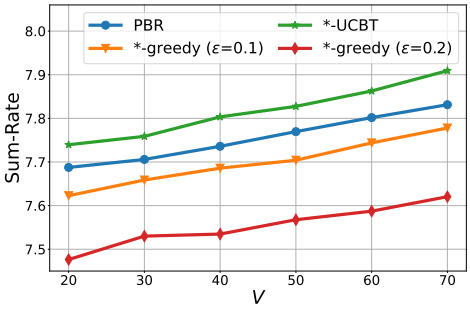

Fig. 11. Comparison of PBR and its variants ($'*'$ stands for PBR).

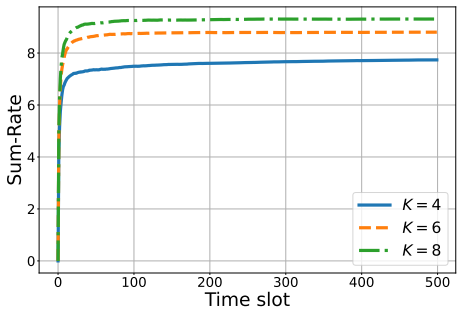

Fig. 9. Sum-rate of PBR with different numbers of users.

the higher the value of N, the higher the sum-rate PBR can achieve. That is because the more precoding matrices PBR selects, the more effective channel gain information it can obtain, leading to a better estimation for the corresponding precoding matrices and hence a lower regret. However, in practice, the choice of the value of N concerns the overhead of sending precoded pilots, which limits the choice of the value of N.

5) Sum-Rate under Different Exploration Strategies: By leveraging the idea of ε-greedy [21] and UCB-Tuned (UCBT) [22], we propose two variants of PBR to investigate the sum-rates under different exploration strategies. The details of the two variants are shown as follows.

- PBR-UCBT: As a variant of PBR, PBR-UCBT replaces the estimate $\tilde{R}_i(t)$ in equation (21) and line 3 of Algorithm 1 with UCBT estimate of R_i for all $i \in \mathcal{J}$, and the rest remains the same as PBR.
- PBR-greedy: PBR-greedy employs the idea of ε-greedy algorithm to explore with a probability of ε in the selection of precoding matrices. Specifically, PBR-greedy differs from PBR in line 3 and line 7 of Algorithm 1. Particularly, it replaces the estimate $\tilde{R}_i(t)$ in line 3 with the empirical mean $\bar{R}_i(t)$ for all $i \in \mathcal{J}$. Then in line 7, with probability ε, it selects N precoding matrices uniformly randomly from the codebook, and with probability $1-\varepsilon$, it selects N precoding matrices according to the solution of problem (24). The rest of PBR-greedy remains the same as PBR.

The comparison between PBR and its variants, PBR-UCBT and PBR-greedy with $\varepsilon \in \{0.1, 0.2\}$, is presented in Figure 11. The results show that compared with ε-greedy variants, PBR and PBR-UCBT can achieve relatively high sum-rates due to their adaptive exploration strategies. The superiority of PBR-UCBT is given by its consideration of the variance of rewards in the estimation, which also makes it of higher complexity and harder to be analyzed.

6) Comparison between PBR and Baselines: To further demonstrate the effectiveness of PBR, we compare it with two baselines: Fast-UCB [14] and MEXP3 [23]. The details of the two baselines are as follows.

- Fast-UCB: Fast-UCB leverages the idea of UCB for reward estimation. Besides the upper bound, it further considers the lower bound of the estimation. Specifically, in each

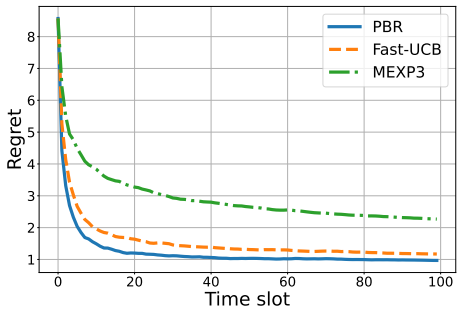

Fig. 12. Comparison of PBR and baselines in terms of regret.

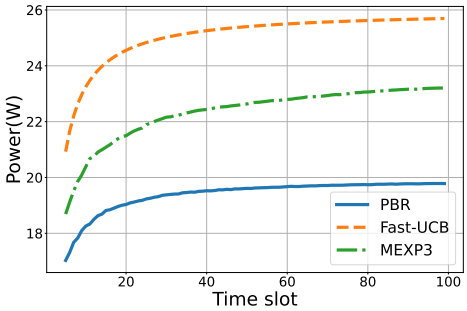

Fig. 13. Comparison of PBR and baselines in terms of power consumption.

time slot, it will eliminate precoding matrices whose upper bounds are less than the lower bound of the precoding matrix with the greatest reward estimation, as those are unlikely to be the optimal selection. By doing so, it can dwindle the codebook size over time and hence improve the learning efficiency.

- MEXP3: MEXP3 is a modified version of EXP3 [20]. In each time slot, the selection of precoding matrices is sampled from a distribution where the probability of a precoding matrix being selected depends on its cumulative rewards plus a normalized constant. According to the distribution, MEXP3 gives all the precoding matrices a chance to be selected. The more accumulated rewards of a precoding matrix, the higher the probability of it being selected. As for those precoding matrices of low accumulated rewards, they also have a chance to be selected due to the intrinsic randomness of MEXP3 for exploration.

The comparison results are shown in Figure 12 and Figure 13, where the value of V and the power constraint p for PBR are $V = 60$ and $p = 20$, respectively.

In Figure 12, we can see that compared with PBR and Fast-UCB, MEXP3 achieves the highest regret and lowest convergence rate, which is incurred by its intrinsic randomness for exploration. Besides, notice that the regret of Fast-UCB is higher than that of PBR. A reason for that may lie in the prune of Fast-UCB: it must be conducted meticulously otherwise it is likely to eliminate precoding matrices with high potential but perform poorly in a couple of time slots.

Figure 13 illustrates the power consumption of the three schemes. Intuitively, a higher transmit power is more likely lead to a higher sum-rate and a lower regret. Given that both Fast-UCB and MEXP3 take no consideration of power consumption specially, the power consumption relation of the two has no exception, *i.e.*, though Fast-UCB achieves a lower regret than MEXP3, its power consumption is much higher than that of MEXP3. However, compared with Fast-UCB and MEXP3, PBR shows its great superiority in such a tradeoff due to the effective integration of online learning and online control. Specifically, by solving problem (24), PBR is able to find precoding matrices with high rewards while ensuring the long-term power constraint.

VI. Conclusion

In this paper, we proposed a novel beamforming design framework based on bandit learning methods and Lyapunov optimization techniques to adaptively learn the best precoding action for a RSMA-aided downlink massive MIMO without explicit CSI feedback. To maximize the ergodic sum-rate while ensuring a long-term transmit power constraint, we propose a precoder selection with bandit learning algorithm for RSMA (PBR). Our theoretical analysis shows that PBR achieves a sub-linear regret bound with a long-term power constraint guarantee. Through experimental results we not only verify our theoretical analysis but also demonstrate the outperformance of PBR in terms of sum-rate and power consumption compared with the conventional transmission schemes without using RSMA.

References

[1] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *arXiv preprint arXiv:2201.03192*, 2022.

[2] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–54, 2018.

[3] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4611–4624, 2016.

[4] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Transactions on Wireless Communications*, Early Access, 2021.

[5] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4847–4861, 2016.

[6] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting for multi-user multi-antenna wireless information and power transfer," in *Proceedings of SPAWC Workshops*, 2019.

[7] J. Park, J. Choi, N. Lee, W. Shin, and H. V. Poor, "Sum spectral efficiency optimization for rate splitting in downlink MU-MISO: A generalized power iteration approach," in *Proceedings of WCNC Workshops*, 2021.

[8] Y. Mao and B. Clerckx, "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6775–6791, 2020.

[9] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8754–8770, 2019.

[10] L. Yin, B. Clerckx, and Y. Mao, "Rate-splitting multiple access for multi-antenna broadcast channels with statistical CSIT," in *Proceedings of IEEE WCNC Workshops*, 2021.
[11] N. Q. Hieu, D. T. Hoang, D. Niyato, and D. I. Kim, "Optimal power allocation for rate splitting communications with deep reinforcement learning," *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2820–2823, 2021.
[12] R. Chataut and R. Akl, "Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction," *Sensors*, vol. 20, no. 10, 2020.
[13] J. Liu, C.-H. R. Lin, Y.-C. Hu, and P. K. Donta, "Joint beamforming, power allocation, and splitting control for SWIPT-enabled IoT networks with deep reinforcement learning and game theory," *Sensors*, vol. 22, no. 6, 2022.
[14] D. Kim, H. V. Poor, and N. Lee, "Online learning to precode for FDD massive MIMO systems," in *Proceedings of IEEE GLOBECOM Workshops*, 2020.
[15] K. Xu, F.-C. Zheng, P. Cao, H. Xu, X. Zhu, and X. Xiong, "DNN-aided codebook based beamforming for FDD millimeter-wave massive MIMO systems under multipath," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 437–452, 2022.
[16] J. Wang, S. Wang, Y. Mao, and Z. Shao, "Online learning-based beamforming for rate-splitting multiple access: A constrained bandit approach," ShangheiTech University, Tech. Rep., 2022. [Online]. Available: http://faculty.sist.shanghaitech.edu.cn/faculty/shaozy/rsma.pdf
[17] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.
[18] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," in *Proceedings of IEEE INFOCOM*, 2019.
[19] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proceedings of ICML*, 2013.
[20] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.
[21] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proceedings of ECML*, 2005.
[22] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
[23] I. Chafaa, E. V. Belmega, and M. Debbah, "Exploiting channel sparsity for beam alignment in mmwave systems via exponential learning," in *Proceedings of IEEE ICC Workshops*, 2020.

APPENDIX A
DERIVATIONS OF THE PER-ROUND SUB-PROBLEMS

In this appendix, we derive the per-time-slot sub-problem. Recall that we define the precoder selection decisions PBR as $\boldsymbol{I}(t) = (I_i(t))_{i\in\mathcal{J}}$ and the corresponding vector $\tilde{\boldsymbol{I}}(t) = (\tilde{I}_i(t))_{i\in\mathcal{J}}$ to indicate the index of the optimal precoding matrix.

We first define the Lyapunov function as follows:

$$L(Q(t)) \triangleq \frac{1}{2}(Q(t))^2,$$

where recall that $Q(t)$ is the queue backlogs at the beginning of time slot t.

Next, we apply the Lyapunov *drift-plus-regret* techniques: We first define the difference of Lyapunov function between two consecutive time slots as $\Delta_t \triangleq L(Q(t+1)) - L(Q(t))$, and we have the following:

$$\begin{aligned}\Delta_t =& L(\boldsymbol{Q}(t+1)) - L(\boldsymbol{Q}(t)) = \frac{1}{2}\Big[(Q(t+1))^2 - (Q(t))^2\Big]\\ \overset{(a)}{\leq}& \frac{1}{2}\left\{\Big[Q(t)+P(t)-p\Big]^2 - (Q(t))^2\right\}\\ =& \frac{1}{2}\Big(P(t)-p\Big)^2 + Q(t)\Big(P(t)-p\Big)\\ \overset{(b)}{\leq}& \frac{1}{2}\max\Big\{(p_{\max}-p)^2, p^2\Big\} + Q(t)\Big(P(t)-p\Big)\end{aligned}$$

where inequality (a) holds by the update equations of the virtual queue in (22); (b) holds since $P_i \leq p_{\max}, \forall i \in \mathcal{J}$, and $\sum_{i\in\mathcal{J}} \tilde{I}_i(t) = 1$.

Next, we define constant D as follows:

$$D \triangleq \frac{1}{2}\max\Big\{(p_{\max}-p)^2, p^2\Big\}.$$

In other words, we have the following upper bound of Δ_t:

$$\Delta_t \leq D + Q(t)\Big(P(t)-p\Big).$$

Further, we define the *Lyapunov drift* as the conditional expectation of Δ_t:

$$\begin{aligned}&\Delta L(Q(t)) \triangleq \mathbb{E}\Big[\Delta_t | Q(t)\Big]\\ \leq& D - Q(t)p + Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i\tilde{I}_i(t)|Q(t)\right].\end{aligned} \tag{25}$$

According to [18], we consider a special class of policies called W-only policy. A W-only policy observes the predefined codebook $\mathcal{P}$ and the power constraint p and makes randomized and stationary decisions depending only on them. Consider an optimal W-only policy which makes decision $\boldsymbol{I}^*(t) = \{I_i^*(t)\}_{i\in\mathcal{N}}$ during each time slot t. Recall that we have the following notation about the regret:

$$\mathrm{Reg}(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\max_{i:I_i^*(t)=1} R_i(t) - \max_{i:I_i(t)=1} R_i(t)\right],$$

thus the *per-time-slot regret* $\Delta\mathrm{Reg}(t)$ can be denoted as follows:

$$\Delta\mathrm{Reg}(t) \triangleq \max_{i:I_i^*(t)=1} R_i(t) - \max_{i:I_i(t)=1} R_i(t).$$

In the following, we first define the *drift-plus-regret* term as

$$\Delta_V(Q(t)) \triangleq \Delta L(Q(t)) + V\mathbb{E}\Big[\Delta\mathrm{Reg}(t)|Q(t)\Big],$$

where V is a positive tunable parameter and we have:

$$\begin{aligned}\Delta_V(Q(t)) \leq\ & D - Q(t)p + Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i\tilde{I}_i(t)|Q(t)\right]\\ &+ V\mathbb{E}\left[\max_{i:I_i^*(t)=1} R_i(t) - \max_{i:I_i(t)=1} R_i(t)|Q(t)\right].\end{aligned} \tag{26}$$

In the following, we derive the per-time-slot sub-problems based on the following facts:

$$\begin{aligned}\max_{i:I_i(t)=1} R_i(t) &= \max_{i:I_i(t)=1}\hat{R}_i(t) + \max_{i:I_i(t)=1} R_i(t) - \max_{i:I_i(t)=1}\hat{R}_i(t)\\ &\geq \max_{i:I_i(t)=1}\hat{R}_i(t) + r_{\min} - r_{\max},\end{aligned}$$

where recall that $\hat{R}_i(t)$ is the upper confidence bound of estimation of sum-rate for precoding matrix i till time slot t.

We then have the upper bound of the drift-plus-regret term:

$$\begin{aligned}&\Delta_V(Q(t)) = \Delta L(Q(t)) + V\mathbb{E}\Big[\Delta\mathrm{Reg}(t)|Q(t)\Big]\\ \leq& D + V(r_{\max}-r_{\min}) - Q(t)p\\ &+ Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{N}} P_i\tilde{I}_i(t)|Q(t)\right]\\ &+ V\mathbb{E}\left[\max_{i:I_i^*(t)=1} R_i(t) - \max_{i:I_i(t)=1}\hat{R}_i(t)|Q(t)\right]\\ =& D + V(r_{\max}-r_{\min}) - Q(t)b\\ &+ V\mathbb{E}\left[\max_{i:I_i^*(t)=1} R_i(t)|Q(t)\right]\\ &- \mathbb{E}\left[V\max_{i:I_i(t)=1}\hat{R}_i(t) - \sum_{i\in\mathcal{J}}\Big(Q(t)P_i\Big)\tilde{I}_i(t)|Q(t)\right],\end{aligned} \tag{27}$$

Recall that, we have

$$w_i(t) = Q(t)P_i.$$

Therefore, to minimize the upper bound of the drift-plus-regret term, we consider the following per-time-slot sub-problems during each time slot t:

$$\begin{aligned}\underset{\boldsymbol{I}(t)}{\text{maximize}}\ & V\max_{i:I_i(t)=1}\hat{R}_i(t) - \sum_{i\in\mathcal{N}} w_i(t)\tilde{I}_i(t)\\ \text{subject to}\ & \sum_{i\in\mathcal{J}} I_i(t) = N, I_i(t)\in\{0,1\}, \forall i\in\mathcal{J},\end{aligned} \tag{28}$$

where recall that $\tilde{I}_i(t) = 1$ if $i = \tilde{i}(t) = \mathrm{argmax}_{j\in\mathcal{J}} R_j(t)I_j(t)$. In this way, we can solve the original problem approximately optimally by solving a series of per-time-slot sub-problems.

APPENDIX B
PROOFS OF QUEUE STABILITY

In this appendix, we prove the queue stability of PBR. Recall that we define the precoder selection decisions under PBR as $\boldsymbol{I}(t) = (I_i(t))_{i\in\mathcal{J}}$ and the corresponding vector $\tilde{\boldsymbol{I}}(t) = (\tilde{I}_i(t))_{i\in\mathcal{J}}$ to indicate the index of the optimal precoding matrix.

According to [18], there exists a feasible random policy selecting precoding matrices during time slot t while subjecting to the long-term energy constraint. The precoder selection decision of such a policy during time slot t is denoted by $\boldsymbol{I}^{\boldsymbol{\epsilon}}(t) = (I_i^{\epsilon}(t))_{i\in\mathcal{J}}$ and the corresponding vector to indicate the index of the optimal precoding matrix is defined as $\tilde{\boldsymbol{I}}^{\boldsymbol{\epsilon}}(t) = (\tilde{I}_i^{\epsilon}(t))_{i\in\mathcal{J}}$. Besides, for such a policy, there exist a positive constant ϵ such that for each time slot t:

$$\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i \tilde{I}_i^{\epsilon}(t)\right] + \epsilon \leq p. \tag{29}$$

Therefore, we have the following based on (25):

$$\begin{aligned}
&\Delta L(Q(t)) = \mathbb{E}\Big[\Delta_t | Q(t)\Big] \\
\leq& D - Q(t)p + Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{N}} P_i \tilde{I}_i(t) | Q(t)\right] \\
\leq& D + V r_{\max} - Q(t)p \\
&- \mathbb{E}\Big[V \max_{i:I_i(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t) | Q(t)\Big] \\
\leq& D + V r_{\max} - Q(t)p \\
&- \mathbb{E}\Big[V \max_{i:I_i^{\epsilon}(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i^{\epsilon}(t) | Q(t)\Big] \\
\leq& D + V r_{\max} + Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i \tilde{I}_i^{\epsilon}(t) - p | Q(t)\right] \\
\leq& D + V r_{\max} - \epsilon Q(t).
\end{aligned}$$

We summarize the derivations as follows:

$$\mathbb{E}\Big[L(Q(t+1)) - L(Q(t)) | Q(t)\Big] \leq D + V r_{\max} - \epsilon Q(t).$$

Taking expectation at both sides of above inequality and summing over time slots $\{0, \ldots, T'-1\}$ gives the following:

$$\mathbb{E}\Big[L(Q(T'))\Big] - \mathbb{E}\Big[L(Q(0))\Big] \leq T'D + T'V r_{\max} - \epsilon \sum_{t=0}^{T'-1} \mathbb{E}\left[Q(t)\right].$$

Since $L(Q(T')) \geq 0$ and $L(Q(0)) = 0$, we have the following by rearranging the above inequality:

$$\frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}\left[Q(t)\right] \leq \frac{D + V r_{\max}}{\epsilon}.$$

By taking the lim-sup of the above inequality as $T' \to \infty$, we obtain the following:

$$\limsup_{T'\to\infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}\left[Q(t)\right] \leq \frac{D + V r_{\max}}{\epsilon} < \infty,$$

where $D = \frac{1}{2}\max\Big\{(p_{\max} - p)^2, p^2\Big\}$; V is a positive tunable parameter of finite value.

Appendix C
Derivations of Regret Bound

In this appendix, we derive the regret bound for PBR. Recall that we define the precoder selection decisions under PBR as $\boldsymbol{I}(t) = (I_i(t))_{i\in\mathcal{J}}$ and the corresponding vector $\tilde{\boldsymbol{I}}(t) = (\tilde{I}_i(t))_{i\in\mathcal{J}}$ to indicate the index of the optimal precoding matrix.

We first derive a general upper bound of the time-averaged regret that does not depend on specific characteristics of the algorithm deigns Recall that the randomized and stationary precoder selection decision during each time slot t under the optimal policy is denoted by $\boldsymbol{I}^{*}(t) = (I_i^{*}(t))_{i\in\mathcal{J}}$ and the corresponding vector $\tilde{\boldsymbol{I}}^{*}(t) = (\tilde{I}_i^{*}(t))_{i\in\mathcal{J}}$ to indicate the index of the optimal precoding matrix.

Therefore, it follows that during each time slot t

$$\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i \tilde{I}_i^{*}(t)\right] \leq p. \tag{30}$$

Based on (26), we have the following:

$$\begin{aligned}
\Delta_V(\boldsymbol{Q}(t)) \leq& D - Q(t)p + Q(t)\mathbb{E}\left[\sum_{i\in\mathcal{J}} P_i \tilde{I}_i(t) | Q(t)\right] \\
&+ V\mathbb{E}\left[\max_{i:I_i^{*}(t)=1} R_i(t) - \max_{i:I_i(t)=1} R_i(t) | Q(t)\right] \\
\leq& D + \mathbb{E}\left[V \max_{i:I_i^{*}(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i^{*}(t) | Q(t)\right] \\
&- \mathbb{E}\left[V \max_{i:I_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t) | Q(t)\right].
\end{aligned}$$

For convenience, we define

$$\begin{aligned}
Z_1(t) =& \left[V \max_{i:I_i^{*}(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i^{*}(t)\right] \\
&- \left[V \max_{i:I_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t)\right].
\end{aligned}$$

Therefore, we have

$$\Delta_V(Q(t)) \leq D + \mathbb{E}\Big[Z_1(t) | Q(t)\Big].$$

Taking expectation and summing over $\{0, \ldots, T-1\}$ of the above equation gives the following:

$$\sum_{t=0}^{T-1} \mathbb{E}\Big[\Delta_t\Big] + V\sum_{t=0}^{T-1} \mathbb{E}\Big[\Delta\text{Reg}(t)\Big] \leq DT + \sum_{t=0}^{T-1} \mathbb{E}\Big[Z_1(t)\Big],$$

and we further divide both sides of the inequality by TV:

$$\begin{aligned}
&\frac{1}{TV}\Big(\mathbb{E}\Big[L(Q(T))\Big] - \mathbb{E}\Big[L(Q(0))\Big]\Big) + \text{Reg}(T) \\
&\leq \frac{D}{V} + \frac{1}{TV}\sum_{t=0}^{T-1} \mathbb{E}\Big[Z_1(t)\Big].
\end{aligned}$$

Since $L(Q(T)) \geq 0$ and $L(Q(0)) = 0$, we have

$$\text{Reg}(T) \leq \frac{D}{V} + \frac{1}{TV}\sum_{t=0}^{T-1} \mathbb{E}\Big[Z_1(t)\Big].$$

Given the above derivation, we can derive upper bounds of regret by bounding $Z_1(t)$ from above. To this end, we assume a policy under which the agent makes independent decisions by solving the following problem (31) optimally during each time slot t. Accordingly, we denote precoder selection decision conducted under such a policy by $\boldsymbol{I'}(t) = \{I'_i(t)\}_{i\in\mathcal{J}}$ and its corresponding vector to indicate the index of the optimal precoding matrix is $\tilde{\boldsymbol{I}}'(t) = (\tilde{I'}_i(t))_{i\in\mathcal{J}}$.

$$\begin{aligned} &\underset{\boldsymbol{I}(t)}{\text{maximize}}\ V \max_{i:I_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t) \\ &\text{subject to}\ \sum_{i\in\mathcal{J}} I_i(t) = N, I_i(t) \in \{0,1\}, \forall i \in \mathcal{J}. \end{aligned} \tag{31}$$

To proceed, we derive upper bounds of $Z_1(t)$. From problem (31), we observe that

$$\begin{aligned} &V \max_{i:I_i^*(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i^*(t) \\ &\le V \max_{i:I'_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}'_i(t). \end{aligned} \tag{32}$$

Similarly, recall the per-time-slot sub-problem (28) we have defined, we have

$$\begin{aligned} &V \max_{i:I'_i(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}'_i(t) \\ &\le V \max_{i:I_i(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t). \end{aligned} \tag{33}$$

Based on (32) and (33), we obtain

$$\begin{aligned} Z_1(t) =& \left[V \max_{i:I_i^*(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i^*(t)\right] \\ &- \left[V \max_{i:I_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t)\right] \\ \le& \left[V \max_{i:I'_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}'_i(t)\right] \\ &- \left[V \max_{i:I_i(t)=1} R_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t)\right] \\ &+ \left[V \max_{i:I_i(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}_i(t)\right] \\ &- \left[V \max_{i:I'_i(t)=1} \hat{R}_i(t) - \sum_{i\in\mathcal{J}} w_i(t)\tilde{I}'_i(t)\right] \\ =& V\left[\max_{i:I_i(t)=1} \hat{R}_i(t) - \max_{i:I_i(t)=1} R_i(t)\right] \\ &+ V\left[\max_{i:I'_i(t)=1} R_i(t) - \max_{i:I'_i(t)=1} \hat{R}_i(t)\right]. \end{aligned}$$

To simplify the above inequality, we define

$$\hat{i} \triangleq \arg\max_{i:I_i(t)=1} \hat{R}_i(t), i' \triangleq \arg\max_{i:I'_i(t)=1} R_i(t).$$

Since $\max_{i:I_i(t)=1} R_i(t) \ge R_{\hat{i}}(t)$ and $\max_{i:I'_i(t)=1} \hat{R}_i(t) \ge \hat{R}_{i'}(t)$, we gain an upper bound of $Z_1(t)$ as follows:

$$Z_1(t) \le V\Big[\hat{R}_{\hat{i}}(t) - R_{\hat{i}}(t)\Big] + V\Big[R_{i'}(t) - \hat{R}_{i'}(t)\Big].$$

Note that both $\hat{\mu}_{i'}(t)$ and $\hat{\mu}_{\hat{i}}(t)$ only make use of information until time slot $(t-1)$ by definition and are not coupled with randomness during time slot t, specifically, $R_i(t)$, $\boldsymbol{I}(t)$ and $\boldsymbol{I'}(t)$. Thus we have

$$\mathbb{E}\Big[Z_1(t)|\hat{i}, i'\Big] \le V\mathbb{E}\Big[\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}|\hat{i}\Big] + V\mathbb{E}\Big[\mu_{i'} - \hat{R}_{i'}(t)|i'\Big],$$

and we define

$$Z_2(t) \triangleq \mathbb{E}\Big[\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}|\hat{i}\Big], Z_3(t) \triangleq \mathbb{E}\Big[\mu_{i'} - \hat{R}_{i'}(t)|i'\Big].$$

To derive an upper bound of $Z_1(t)$, it is necessary to bound both $Z_2(t)$ and $Z_3(t)$.

Bounding $\sum_{t=0}^{T-1} Z_2(t)$. We give an upper bound of $\sum_{t=0}^{T-1} Z_2(t)$ by defining the event $F(t) = \{\hat{R}_{\hat{i}}(t) \ge \mu_{\hat{i}}|\hat{i}\}$ and its complementary event as $F^c(t)$. Then we have

$$\begin{aligned} Z_2(t) =& \mathbb{E}\left[\Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\Big(\mathbb{I}\{F(t)\} + \mathbb{I}\{F^c(t)\}\Big)|\hat{i}\right] \\ \overset{(a)}{\le}& \mathbb{E}\left[\Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\}|\hat{i}\right], \end{aligned}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function which takes value 1 indicating the occurrence of event and 0 otherwise, and (a) holds since the occurrence of event $F^c(t)$ implies that $\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}} < 0$. For convenience, we define

$$J^{(1)}(t) = (\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}})\mathbb{I}\{F(t)\}.$$

We can further define event $G(t) = \{\bar{R}_{\hat{i}}(t) - \mu_{\hat{i}} > \lambda\gamma_{\hat{i}}(t)|\hat{i}\}$ and its complementary event as $G^c(t)$ where $\gamma_i(t) = \sqrt{\log t / z_i(t)}$. And we have the following derivation:

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=0}^{T-1} J^{(1)}(t)|\hat{i}\right] \\ \le& r_{\max} - r_{\min} + \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\}|\hat{i}\right] \\ =& r_{\max} - r_{\min} + \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\cap G(t)\}|\hat{i}\right] \\ &+ \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\cap G^c(t)\}|\hat{i}\right], \end{aligned}$$

where $t_{\hat{i}}^{(1)}$ is the first time slot each precoding matrix $\hat{i}$ is selected. For the last two terms of the above inequality, we denote them as $J^{(2)}(t)$ and $J^{(3)}(t)$ and bound them separately. Specifically,

$$\begin{aligned} J^{(2)}(t) \triangleq& (\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}})\mathbb{I}\{F(t)\cap G(t)\}, \\ J^{(3)}(t) \triangleq& (\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}})\mathbb{I}\{F(t)\cap G^c(t)\}. \end{aligned}$$

Bounding $\mathbb{E}\Big[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(2)}(t)|\hat{i}\Big]$. Provided that the $G(t)$ occurs, either $\hat{R}_{\hat{i}}(t) = \bar{R}_{\hat{i}} + \lambda\gamma_{\hat{i}}(t)$ or $\hat{R}_{\hat{i}}(t) = r_{\max}$ gives $\hat{R}_{\hat{i}}(t) > \mu_{\hat{i}}$. *i.e.*, the event $F(t)$ occurs. Accordingly, we have $G(t) \subset F(t)$. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(2)}(t)|\hat{i}\right] \\
=&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\cap G(t)\}|\hat{i}\right] \\
=&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{G(t)\}|\hat{i}\right] \\
\overset{(a)}{\leq}&(r_{\max}-r_{\min}) \sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} Pr\Big\{\bar{R}_{\hat{i}}(t) - \mu_{\hat{i}} > \lambda\gamma_{\hat{i}}(t)|\hat{i}\Big\} \\
\leq&(r_{\max}-r_{\min}) \\
&\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \sum_{i:I_i(t)=1} Pr\Big\{\bar{R}_j(t) - \mu_j > \lambda\gamma_j(t)|j=i\Big\} \\
\overset{(b)}{\leq}&(r_{\max}-r_{\min}) \sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \sum_{i:I_i(t)=1} \exp\left(-2\lambda^2 z_i(t)\frac{\log t}{z_i(t)}\right) \\
\leq&N(r_{\max}-r_{\min})\sum_{t=1}^{\infty} t^{-2\lambda^2} = N(r_{\max}-r_{\min})H_\lambda
\end{aligned}
$$

where (a) holds since $\hat{R}_{\hat{i}}(t) \in [0, r_{\max}], \mu_{\hat{i}} \in [r_{\min}, r_{\max}]$; (b) is true due to the Hoeffding's inequality, and $H_\lambda \triangleq \sum_{t=1}^{\infty} t^{-2\lambda^2}$ for simplicity.

Bounding $\mathbb{E}\Big[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\Big]$. Recall that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\right] \\
=&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} \Big(\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}\Big)\mathbb{I}\{F(t)\cap G^c(t)\}|\hat{i}\right].
\end{aligned}
$$

Given that the event $G^c(t)$ occurs, we have the following:

$$
\begin{aligned}
&\hat{R}_{\hat{i}}(t) = \min\Big\{\bar{R}_{\hat{i}}(t) + \lambda\gamma_{\hat{i}}(t), r_{\max}\Big\} \leq \bar{R}_{\hat{i}}(t) + \lambda\gamma_{\hat{i}}(t) \\
&\Rightarrow \hat{R}_{\hat{i}}(t) - \mu_{\hat{i}} = \hat{R}_{\hat{i}}(t) - \bar{R}_{\hat{i}}(t) + \bar{R}_{\hat{i}}(t) - \mu_{\hat{i}} \leq 2\lambda\gamma_{\hat{i}}(t).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\right] \\
\leq&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} 2\lambda\gamma_{\hat{i}}(t)\mathbb{I}\{F(t)\cap G^c(t)\}|\hat{i}\right] \\
\leq&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} 2\lambda\gamma_{\hat{i}}(t)|\hat{i}\right].
\end{aligned}
$$

We then further extend the bound as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\right] \leq \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} 2\lambda\gamma_{\hat{i}}(t)|\hat{i}\right] \\
\leq&\mathbb{E}\left[\sum_{z=2}^{z_k(T-1)} 2\lambda\gamma_{\hat{i}}(t_{\hat{i}}^{(p)})|\hat{i}\right] = 2\lambda\mathbb{E}\left[\sum_{z=2}^{z_k(T-1)} \sqrt{\frac{\log t}{z_{\hat{i}}(t_{\hat{i}}^{(p)})}}|\hat{i}\right] \\
\leq&2\lambda\mathbb{E}\left[\sqrt{\log T}\Big(\int_2^{z_{\hat{i}}(T-1)} \frac{1}{\sqrt{z}}dz + 1\Big)|\hat{i}\right] \\
\leq&4\lambda\sqrt{\log T}\mathbb{E}\left[\sqrt{z_{\hat{i}}(T-1)}|\hat{i}\right] \\
\leq&4\lambda\sqrt{\log T}\mathbb{E}\left[\sum_{i\in\mathcal{J}}\sqrt{z_i(T-1)}\right]
\end{aligned}
$$

where $z_{\hat{i}}(t)$ denote the number of time slots that precoding matrix $\hat{i}$ has been selected at the beginning of time slot t.

In summary, by combining the upper bound of $\mathbb{E}\Big[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(2)}(t)|\hat{i}\Big]$ and $\mathbb{E}\Big[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\Big]$, we have the following:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} Z_2(t) \\
\leq&r_{\max} - r_{\min} + \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(2)}(t)|\hat{i}\right] + \mathbb{E}\left[\sum_{t=t_{\hat{i}}^{(1)}+1}^{T-1} J^{(3)}(t)|\hat{i}\right] \\
\leq&r_{\max} - r_{\min} + N(r_{\max}-r_{\min})H_\lambda \\
&+ 4\lambda\sqrt{\log T}\mathbb{E}\left[\sum_{i\in\mathcal{J}}\sqrt{z_i(T-1)}\right] \\
\leq&r_{\max} - r_{\min} + N(r_{\max}-r_{\min})H_\lambda \\
&+ 4\lambda\sqrt{\log T}\mathbb{E}\left[\sqrt{\sum_{i\in\mathcal{N}} z_i(T-1)}\right] \\
\leq&r_{\max} - r_{\min} + N(r_{\max}-r_{\min})H_\lambda + 4\lambda\sqrt{NT\log T}.
\end{aligned}
$$

Bounding $\sum_{t=0}^{T-1} Z_3(t)$. We give an upper bound of $\sum_{t=0}^{T-1} Z_3(t)$ by defining the event $W(t) = \{\hat{R}_{i'} \geq \mu_{i'}|i'\}$ and its complementary event as $W_c(t)$. Recall that

$$
Z_3(t) = \mathbb{E}\Big[\mu_{i'} - \hat{R}_{i'}|i'\Big],
$$

and we define

$$J^{(4)}(t) \triangleq (\mu_{i'} - \hat{R}_{i'})\mathbb{I}\{W^c(t)\}. \tag{34}$$

On one hand, when $t \leq t_{i'}^{(1)}$, we have $\hat{R}_{i'}(t) = r_{\max}$, and event $W^c(t)$ will not occur, hence $J^{(4)}(t) = 0$. On the other hand, when $t \geq t_{i'}^{(1)}+1$, if event $W^c(t)$ occurs, we have $\hat{R}_{i'}(t) \leq \mu_{i'} \leq r_{\max}$, which implies that $\hat{R}_{i'} = \bar{R}_{i'} + \lambda\sqrt{\log t / z_{i'}(t)}$ and thus $\mu_{i'} \geq \bar{R}_{i'}(t) + \lambda\sqrt{\log t / z_i(t)}$, we have the following:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} J^{(4)}(t)|i'\right] = \mathbb{E}\left[\sum_{t=t_{i'}^{(1)}+1}^{T-1} J^{(4)}(t)|i'\right] \\
\leq &\mathbb{E}\left[\sum_{t=t_{i'}^{(1)}+1}^{T-1} (\mu_{i'} - \hat{R}_{i'})\mathbb{I}\{F^c(t)\}|i'\right] \\
\overset{(a)}{\leq} &\mathbb{E}\left[\sum_{t=t_{i'}^{(1)}+1}^{T-1} \mathbb{I}\{F^c(t)\}|i'\right] \\
\leq &\sum_{t=t_{i'}^{(1)}+1}^{T-1} \sum_{i:I'(t)=1} Pr\left\{\mu_j \geq \bar{R}_j(t) + \lambda\sqrt{\frac{\log t}{z_j(t)}}\Big| j = i\right\} \\
\leq &N\sum_{t=1}^{\infty} t^{-2\lambda^2} = NH_\lambda,
\end{aligned}$$

where (a) holds due to the occurrence of event $W^c(t)$.

Therefore, we can have an upper bound of $\sum_{t=0}^{T-1} Z_3(t)$ as follows:

$$\sum_{t=0}^{T-1} Z_3(t) = \mathbb{E}\left[\sum_{t=t_{i'}^{(1)}+1}^{T-1} J^{(4)}(t)|i'\right] \leq NH_\lambda.$$

Recall that

$$\mathbb{E}\Big[Z_1(t)|\hat{i}, i'\Big] \leq V\mathbb{E}\Big[\hat{R}_{\hat{i}}(t) - \mu_{\hat{i}}|\hat{i}\Big] + V\mathbb{E}\Big[\mu_{i'} - \hat{R}_{i'}(t)|i'\Big],$$

and by taking expectation at both sides of the above inequality, we have the following:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\Big[Z_1(t)\Big] \leq &V(r_{\max} - r_{\min} + (r_{\max} - r_{\min} + 1)NH_\lambda) \\
&+ 4\lambda V\sqrt{NT\log T}.
\end{aligned}$$

Finally, since $\text{Reg}(T) \leq \frac{D}{V} + \frac{1}{TV}\sum_{t=0}^{T-1}\mathbb{E}\Big[Z_1(t)\Big]$, we have the following upper bound of the time-averaged regret bound:

$$\begin{aligned}
\text{Reg}(T) \leq &\frac{D}{V} + \frac{(r_{\max} - r_{\min} + (r_{\max} - r_{\min} + 1)NH_\lambda)}{T} \\
&+ 4\lambda\sqrt{\frac{N\log T}{T}},
\end{aligned}$$

where $D = \frac{1}{2}\max\Big\{(p_{\max} - p)^2, p^2\Big\}$; $H_\lambda = \sum_{t=1}^{\infty} t^{-2\lambda^2}$; V is a positive tunable parameter for balancing the trade-off between regret minimization and the time-averaged constraint satisfaction.