On Stochastic Feedback Control for Multi-antenna Beamforming: Formulation and Low-Complexity Algorithms

Sun Sun, *Student Member, IEEE*, Min Dong, *Senior Member, IEEE*, and Ben Liang, *Senior Member, IEEE*

***Abstract*—Based on the Gauss-Markov channel model, we investigate the stochastic feedback control for transmit beamforming in multiple-input-single-output (MISO) systems, and design practical implementation algorithms leveraging techniques in dynamic programming and reinforcement learning. We first validate the Markov Decision Process (MDP) formulation of the underlying feedback control problem with a $4R$-variable ($4R$-V) state, where R is the number of the transmit antennas. Due to the high complexity of finding an optimal feedback policy under the $4R$-V state, we consider a reduced 2-V state. As opposed to a previous study that assumes the feedback problem under such a 2-V state remaining an MDP formulation, our analysis indicates that the underlying problem is no longer an MDP. Nonetheless, the approximation as an MDP is shown to be justifiable and efficient. Based on the quantized 2-V state and the MDP approximation, we propose practical implementation algorithms for feedback control with unknown state transition probabilities. In particular, we provide model-based off-line and on-line learning algorithms, as well as a model-free learning algorithm. We investigate and compare these algorithms through extensive simulations, and provide their efficiency analysis. According to these results, the application rule of these algorithms is established under both statistically stable and unstable channels.**

***Index Terms*—beamforming, stochastic feedback control, implementation algorithms, reinforcement learning.**

I. Introduction

Multi-antenna transmit beamforming is an effective physical-layer technique that can improve transmission rate and reliability over a wireless link in a slow fading environment [1]. The actual beamforming gain achieved depends on feedback quality, which controls the accuracy of the beamforming vector used at the transmitter. Thus, to maximally realize the beamforming potential, one key problem of a practical system is to design an efficient feedback strategy to obtain the beamforming vector information at the transmitter. As feedback incurs overhead and thus rate loss to the system, the net gain in terms of the system throughput is the result of both rate gain due to beamforming and rate loss due to feedback.

Channel temporal correlation introduces an additional time dimension for the feedback strategy design. In the literature, beamforming feedback design typically focuses on the quantization techniques of the channel or beamforming vector, *i.e.,* codebook design [2], [3]. The temporal dimension is addressed either assuming block fading channel model with independent fade changing from block to block, or through simple periodic feedback without actively exploring the channel temporal correlation. Recently, a stochastic feedback control approach is proposed for the beamforming feedback design by formulating the problem as a Markov Decision Process (MDP) problem [4]. In this approach, the feedback decision takes into account both the channel temporal correlation and the feedback cost, with the goal of maximizing the system net throughput. Stochastic feedback control is attractive as the feedback decision is dynamic, which potentially could improve the feedback efficiency. Gains of the stochastic feedback over the periodic one are demonstrated in [4] through extensive simulations. In this paper, we focus on analyzing the stochastic feedback control of beamforming vector for transmit beamforming. Specifically, we investigate the feedback problem from the perspectives of channel temporal correlation modeling, problem space reduction and its optimality, as well as low-complexity algorithms for feedback control.

A. Our Contribution

We consider a multi-input single-output (MISO) transmit beamforming system with R transmit antennas and a single receive antenna. The temporal evolution of the channel gain is modeled by a first-order Gauss-Markov process. By modeling the feedback of the beamforming vector as an on-off decision process, we aim to design a feedback policy to maximize the system net throughput under a given average feedback cost. Our contributions are summarized as follows.

- Using the channel gain vector and available beamforming vector at the transmitter, we construct a $4R$-variable ($4R$-V) system state. Based on this system state, we prove that the underlying feedback control problem is an MDP problem, and for the quantized $4R$-V state, there exits an optimal stationary feedback policy that is Markovian (only depends on the current system state) and deterministic. However, due to the "curse of dimensionality," the complexity of obtaining an optimal policy is prohibitive.
- To make the problem tractable for a practical feedback design, we investigate a reduced 2-V state that consists of the channel power and the achieved beamforming gain. In [4], such a reduced state is claimed to incur no loss of optimality in the sense that, the feedback control problem under this 2-V state remains an MDP problem. However,

Sun Sun and Ben Liang are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada (email: {ssun, liang}@comm.utoronto.ca).

Min Dong is with the Department of Electrical Computer and Software Engineering, University of Ontario Institute of Technology, Toronto, Canada (email: min.dong@uoit.ca).

through a detailed analysis, we demonstrate that such a reduction is in fact suboptimal, *i.e.,* the underlying feedback control problem is no longer an MDP problem. Nonetheless, the approximation as an MDP is shown to be justifiable and efficient.

- Based on the quantized 2-V state and the MDP approximation, we propose practical feedback control algorithms without requiring the state transition probabilities. In particular, we provide off-line and on-line model-based learning algorithms that can learn the transition probabilities first and then obtain the feedback policy by dynamic programming (DP) methods. We also provide a model-free reinforcement learning algorithm that can directly learn the feedback policy without obtaining the transition probabilities. To the best of our knowledge, this is the first paper that employs reinforcement learning techniques in the stochastic feedback control for transmit beamforming. In addition, we study and compare these algorithms through extensive simulation, and provide their efficiency analysis. Based on these results, the application rule of these algorithms is suggested for both statistically stable and unstable channels.

B. Related Works

There is a large body of literature on beamforming feedback design [2]. Most previous works focus on codebook design for beamforming vector under limited feedback, with emphasis on the quantization performance of a given channel [2], [3], [5], [6]. Approaches such as Grassmannian line packing [3], vector quantization [6], [7], and random vector quantization [5] are proposed and studied. To further improve the feedback efficiency and reduce feedback overhead, opportunistic feedback approach is considered in multiuser transmission systems [8]–[10], where the threshold-type feedback is assumed and derived for various design metrics. However, these results are derived by assuming a block fading channel model with independent fade from block to block. Thus, channel temporal correlation is not actively explored for improving the feedback efficiency. To explore the temporal correlation of fading channels, the channel dynamics are commonly modeled either by continuous Gauss-Markov models [11]–[14], or finite-state Markov models [15]–[17]. With the channel temporal correlation, feedback compression methods for channel state or beamforming vector are investigated in [18]–[20], and differential quantization strategies are proposed in [13], [14], [21], [22]. Among these results, the periodic feedback strategy is assumed, and the effect of feedback cost on the system throughput is not incorporated in the design. The cost due to feedback is studied in the form of resource allocation (*e.g.*, bandwidth and power) between feedback and data transmission in [23], [24].

A stochastic feedback control approach is first considered in [4]. With the objective of maximizing the system net throughput, [4] aims to find an optimal policy for determining when to feed the beamforming vector back to the transmitter. By modeling the channel temporal evolution through a first-order Gauss-Markov process and viewing the feedback control problem as an MDP problem, the authors show that the feedback policy is of the threshold-type. This work is most relevant to our work. The main differences between our work and [4] are as follows: 1) We rigorously examine the validity of MDP formulation for the feedback control of beamforming vector, while in [4], the MDP formulation is directly assumed; 2) we explore and compare various feedback control algorithms for practical use, while in [4], the focus is on proving the threshold-type property of the optimal policy. Although the existence of a threshold is shown, no practical algorithm is provided to determine the threshold for the feedback control.

Under the MDP formulation, feedback control of beamforming vector is essentially a problem of sequential on-off decision making. When the system state transition probabilities are known, the traditional DP methods, such as policy iteration and value iteration, can be employed to find an optimal policy [25]. In reality, however, the transition probabilities are unavailable. Reinforcement learning machinery provides model-free algorithms, by which an optimal policy can be learned directly without the requirement of the transition probabilities. Excellent tutorials on this subject can be found in [26]–[28].

C. Organization and Notation

The remainder of this paper is organized as follows. In Section II, we present the system model and problem formulation. In Section III, we demonstrate the MDP formulation of the feedback control problem based on a $4R$-V state. In Section IV, we investigate the validity of the MDP formulation based on a reduced 2-V state. In Section V, using the quantized 2-V state, we provide four practical learning algorithms for feedback control. Performance of these algorithms is studied and compared in Section VI followed by a discussion on practical feedback implementation in Section VII. Finally, we conclude in Section VIII.

Notation: Denote the conjugate, transpose, and Hermitian of a matrix $\mathbf{A}$ by $\overline{\mathbf{A}}$, $\mathbf{A}^T$, and $\mathbf{A}^H$, respectively; denote $\mathbf{0}_{m,n}$ as an $m \times n$ matrix with all entries zero; denote $\mathbf{I}_n$ as an identity matrix with the dimension $n \times n$; denote $\log(\cdot)$ as the $\log$ function with base 2; denote $\Re(\cdot)$ and $\Im(\cdot)$ as the real part and the imaginary part of the enclosed parameter respectively; denote $\mathcal{R}$ as the set of all real numbers and $\mathcal{R}^+$ as the set of all non-negative real numbers; denote $\mathbf{1}(\cdot)$ as the indicator function, which equals 1 (resp. 0) when the enclosed statement is true (resp. false); denote $\mathbb{E}\ [\cdot]$ as the expectation. Let $\mathbf{g}$ be an $n \times 1$ vector. Then $\|\mathbf{g}\|$ represents the Euclidean norm of $\mathbf{g}$, and $\mathbf{g} \sim \mathcal{CN}(\mathbf{0}_{n,1}, \sigma^2\mathbf{I}_n)$ means that $\mathbf{g}$ is a circular complex Gaussian random vector with mean zero and covariance $\sigma^2\mathbf{I}_n$. For a process $\{n_t\}$, if all its elements n_t are with mean zero and variance σ^2, and the element n_t is uncorrelated with the element $n_{t'}$ for all $t \neq t'$, then we call $\{n_t\}$ a white noise process and denote it by $\{n_t\} \sim WN(0, \sigma^2)$.

II. Channel Model and Problem Statement

A. Channel Model

Consider a MISO system in which the transmitter is equipped with R (≥ 2) antennas and the receiver is equipped

with a single antenna. Assume a discrete-time slow fading channel model. At time slot t, denote $g_{i,t}$ as the channel gain from the i-th transmit antenna to the receiver and $\mathbf{g}_t \triangleq [g_{1,t}, \cdots, g_{R,t}]^T$ as the channel gain vector. Assume that all channel gains are spatially independent but are temporally correlated. The temporal correlation of the channel vector $\mathbf{g}_t$ is captured by the following first-order Gauss-Markov model:

$$\mathbf{g}_{t+1} = \rho \mathbf{g}_t + \mathbf{w}_{t+1}, \quad t = 0, 1, \cdots. \tag{1}$$

In (1), the initial channel gain $\mathbf{g}_0 \sim \mathcal{CN}(\mathbf{0}_{R,1}, \mathbf{I}_R)$. The noise process $\{\mathbf{w}_t\}$ is independent of $\mathbf{g}_0$, and we assume that $\{\mathbf{w}_t\}$ is i.i.d. with $\mathbf{w}_t \sim \mathcal{CN}\left(\mathbf{0}_{R,1}, (1-\rho^2)\mathbf{I}_R\right)$. ρ is the correlation coefficient defined as $\rho \triangleq J_0(2\pi f_D T_s)$, where $J_0(\cdot)$ is the zero-order Bessel function of the first kind, and $f_D T_s$ is the normalized Doppler frequency with f_D being the maximum Doppler frequency shift and T_s the transmitted symbol period. Obviously, a larger value of ρ indicates a higher temporal correlation of the channel gains[1]. Being simple and effective, the channel model (1) is widely used in the literature to characterize the temporal channel correlation, *e.g.*, [11]–[14].

Using Fact 1 below, we have that $\{\mathbf{g}_t\}$ follows a block Markovian model. Furthermore, we can show that the stationary distribution of $\mathbf{g}_t$ is $\mathcal{CN}(\mathbf{0}_{R,1}, \mathbf{I}_R)$.

Fact 1 ([29]): Let $\epsilon_0, \epsilon_1, \epsilon_2, \cdots$ be independent random variables, $X_0 = \epsilon_0$, and $X_{t+1} = \rho X_t + \epsilon_{t+1}, t = 0, 1, \cdots$, where ρ is a real constant. Then $\{X_t\}$ forms a Markov chain.

B. Problem Statement

At time slot t, denote an $R \times 1$ unite complex vector $\mathbf{b}_{u,t}$, $\|\mathbf{b}_{u,t}\| = 1$, as the available beamforming vector at the transmitter. Then, the received signal at the receiver can be expressed as

$$y_t = x_t \mathbf{b}_{u,t}^H \mathbf{g}_t + v_t \tag{2}$$

where x_t is the transmitted signal with the power constraint $\mathbb{E}\left[|x_t|^2\right] = P$ and v_t is the noise at the receiver. We assume that $\{v_t\}$ is i.i.d. with $v_t \sim \mathcal{CN}(0,1)$, and $\{v_t\}$ is independent of $\{\mathbf{g}_t\}$ and $\{x_t\}$. From (2), we can derive that the signal-to-noise ratio (SNR) is $P|\mathbf{b}_{u,t}^H \mathbf{g}_t|^2$ and the data rate is $\log(1 + P|\mathbf{b}_{u,t}^H \mathbf{g}_t|^2)$.

In practical systems the beamforming vector at the transmitter is obtained through feedback. Hence, the obtained beamforming gain is affected by the quality of the beamformer. Assume that each feedback incurs a fixed cost, denoted by c. Then, there is a trade-off between the feedback cost and the data rate provided by beamforming gain. Combining the feedback cost and beamforming gain, we define the reward at time slot t, r_t, to be the net throughput obtained through beamforming, expressed by

$$r_t = \begin{cases} \log(1 + P\|\mathbf{g}_t\|^2) - c, & \text{if feedback} \\ \log(1 + P|\mathbf{b}_{u,t}^H \mathbf{g}_t|^2), & \text{if no feedback.} \end{cases} \tag{3}$$

[1]For example, when $f_D T_s = 0.01$, *i.e.,* when the channel coherence time is approximately 100 times the symbol period, we have $\rho = 0.999013$, which indicates that the successive channel gains are highly correlated.

The expected total discounted reward is given by

$$V = \mathbb{E}\left[\sum_{t=0}^{\infty} \lambda^t r_t\right] \tag{4}$$

where $\lambda \in (0,1)$ is the discount factor and the expectation is taken over the randomness of the channel gain. With the discount factor λ, the future net throughput is considered to be less valuable than the current one, reflecting the time value in practice.

Our objective is to design a feedback control policy to maximize the expected total discounted reward in (4). To simplify the analysis, throughout the paper, we assume that the receiver knows the channel information perfectly, and the feedback is delay-free and error-free. In other words, upon the feedback, the available beamforming vector at the transmitter is given by

$$\mathbf{b}_{u,t} = \mathbf{g}_{u,t} \triangleq \mathbf{g}_t / \|\mathbf{g}_t\|.$$

III. Optimal Feedback Control: An MDP Formulation

To maximize the expected total discounted reward, the optimal feedback policy may be a randomized one that depends on all previous system states and feedback actions. In this section, we will show that, by introducing a $4R$-V state the underlying feedback control problem can be formulated as an MDP problem. With the quantized $4R$-V state, the formulated MDP problem admits an optimal stationary policy that is Markovian (only depends on the current system state) and deterministic. Unfortunately, finding such an optimal policy can be prohibitive.

A. MDP Formulation

We first give the definition of an MDP problem.

Definition 1 ([25]): Define a collection of objects $\{\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathbb{P}(\cdot|s_t, a_t), r_t(s_t, a_t)\}$ as an MDP, where $\mathcal{T}, \mathcal{S}$, and $\mathcal{A}$ are the spaces of decision epoch t, state s_t, and action a_t, respectively, $\mathbb{P}(\cdot|s_t, a_t)$ is the transition probability of the next state given the current state and action, and $r_t(s_t, a_t)$ is the corresponding reward function.

Define

$$z_t \triangleq (s_0, a_0, s_1, a_1, \cdots, s_t)$$

to be the history of the decision process up to time slot t. The key requirement of an MDP problem is that, the transition probabilities and the reward function should depend on the history z_t only through the current state s_t. We now demonstrate that the underlying feedback control problem can be formulated as an MDP problem.

Let the space of decision epochs be $\mathcal{T} = \{0, 1, \cdots\}$, and the space of actions be $\mathcal{A} = \{0, 1\}$, where 0 represents no feedback and 1 represents feedback. The system state at time slot t is constructed as $s_t = (\mathbf{g}_t, \mathbf{b}_{t-1})$, where the second element $\mathbf{b}_{t-1}$ is the unnormalized version (*i.e.,* including the channel power) of the available beamforming vector at the transmitter at time slot $t-1$. Note that in s_t, $\mathbf{b}_{t-1}$ is the information available at time slot t before the feedback control decision a_t is made. Thus, the state at time slot t contains

$\mathbf{b}_{t-1}$, rather than $\mathbf{b}_t$, which is yet unknown. At each time slot t, based on the system state s_t, the receiver decides whether to feed $\mathbf{g}_{u,t}$ back to the transmitter by choosing $a_t \in \mathcal{A}$. To facilitate later analysis, we also introduce a variable $q_t \in \{1, \cdots, t+1\}$ at time slot t. It denotes that the most recent feedback happens q_t time slot(s) ago from time slot t before the feedback decision a_t is made. Then, we have $\mathbf{b}_{t-1} = \mathbf{g}_{t-q_t}$, with $\mathbf{g}_{-1}$ being the initial unnormalized beamforming vector used at the transmitter at time slot 0. Upon the receiver's decision, the available beamforming vector at the transmitter, the reward, and the state at the next time slot are as follows:

$$\begin{aligned} a_t = 1 \Rightarrow & \begin{cases} \mathbf{b}_{u,t} = \mathbf{g}_{u,t} \triangleq \mathbf{g}_t / \|\mathbf{g}_t\| \\ r_t = \log(1 + P\|\mathbf{g}_t\|^2) - c \\ s_{t+1} = (\mathbf{g}_{t+1}, \mathbf{g}_t), \end{cases} \\ a_t = 0 \Rightarrow & \begin{cases} \mathbf{b}_{u,t} = \mathbf{b}_{u,t-1} = \mathbf{g}_{t-q_t} / \|\mathbf{g}_{t-q_t}\| \\ r_t = \log(1 + P|\mathbf{b}_{u,t}^H \mathbf{g}_t|^2) \\ s_{t+1} = (\mathbf{g}_{t+1}, \mathbf{g}_{t-q_t}). \end{cases} \end{aligned} \tag{5}$$

Note that if $a_t = 0$, the beamforming vector at the transmitter is unchanged, *i.e.,* $\mathbf{b}_{u,t} = \mathbf{b}_{u,t-1}$. Since $\mathbf{b}_{u,t}$ is the normalized version of $\mathbf{b}_t$, we have $\mathbf{b}_t = \mathbf{b}_{t-1} = \mathbf{g}_{t-q_t}$.

In s_t, $\mathbf{g}_t$ and $\mathbf{b}_{t-1}$ are complex vectors. Rewrite $s_t = (\Re(\mathbf{g}_t), \Im(\mathbf{g}_t), \Re(\mathbf{b}_{t-1}), \Im(\mathbf{b}_{t-1}))$ by decomposing each complex variable into the real part and the imaginary part. Then, we have $4R$ real scalar variables in s_t, and we call s_t the $4R$*-V state*. The state space is denoted by $\mathcal{S} = \mathcal{R} \times \cdots \times \mathcal{R}$, which is the Cartesian product of $4R$ real spaces.

It is easy to see that the reward function depends on z_t only through s_t. In the lemma below, we show that this is the case for the transition probabilities as well.

Lemma 1: Based on the $4R$-V state $s_t = (\mathbf{g}_t, \mathbf{b}_{t-1})$, we have

$$\mathbb{P}(s_{t+1}|z_t, a_t) = \mathbb{P}(s_{t+1}|s_t, a_t). \tag{6}$$

Proof: First assume that $a_t = 1$. Then, at time slot $t+1$, the receiver will observe $s_{t+1} = (\mathbf{g}_{t+1}, \mathbf{g}_t)$ by (5). The left hand side (LHS) of (6) is given by

$$\begin{aligned} &\mathbb{P}(s_{t+1}|z_t, a_t = 1) \\ &= \mathbb{P}\left(\mathbf{g}_{t+1}, \mathbf{b}_t \middle| \mathbf{g}_0, \mathbf{b}_{-1}, a_0, \mathbf{g}_1, \mathbf{b}_0, a_1, \cdots, \mathbf{g}_t, \mathbf{b}_{t-1}, a_t = 1\right) && (7) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_t\right) \mathbb{P}\big(\mathbf{g}_{t+1} \big| \mathbf{g}_0, \mathbf{g}_{-1}, \mathbf{g}_1, \mathbf{g}_{1-q_1}, \cdots, \mathbf{g}_t, \mathbf{g}_{t-q_t}\big) && (8) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_t\right) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t\right) && (9) \end{aligned}$$

where (7) is based on the definitions of s_t and z_t, (8) follows because once the actions $\{a_\tau\}_{\tau=0}^t$ are given, we can specify $\mathbf{b}_{\tau-1}$ as $\mathbf{g}_{\tau-q_\tau}$, $\forall \tau \in \{1, \cdots, t+1\}$, and (9) follows due to the Markovian property of $\{\mathbf{g}_t\}$.

The right hand side (RHS) of (6) is given by

$$\begin{aligned} \mathbb{P}(s_{t+1}|s_t, a_t = 1) &= \mathbb{P}\left(\mathbf{g}_{t+1}, \mathbf{b}_t \middle| \mathbf{g}_t, \mathbf{b}_{t-1}, a_t = 1\right) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_t\right) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t, \mathbf{b}_{t-1}\right) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_t\right) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t\right). \end{aligned}$$

Therefore, $\mathbb{P}(s_{t+1}|z_t, a_t = 1) = \mathbb{P}(s_{t+1}|s_t, a_t = 1)$.

Next assume that $a_t = 0$. Then, at time slot $t+1$, the receiver will observe $s_{t+1} = (\mathbf{g}_{t+1}, \mathbf{g}_{t-q_t})$. The LHS of (6) is given by

$$\begin{aligned} &\mathbb{P}(s_{t+1}|z_t, a_t = 0) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_{t-q_t}\right) \mathbb{P}\big(\mathbf{g}_{t+1} \big| \mathbf{g}_0, \mathbf{g}_{-1}, \mathbf{g}_1, \mathbf{g}_{1-q_1}, \cdots, \mathbf{g}_t, \mathbf{g}_{t-q_t}\big) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_{t-q_t}\right) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t\right), \end{aligned}$$

and the RHS of (6) is given by

$$\begin{aligned} \mathbb{P}(s_{t+1}|s_t, a_t = 0) &= \mathbf{1}\big(\mathbf{b}_t = \mathbf{g}_{t-q_t}\big) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t, \mathbf{g}_{t-q_t}\right) \\ &= \mathbf{1}\left(\mathbf{b}_t = \mathbf{g}_{t-q_t}\right) \mathbb{P}\left(\mathbf{g}_{t+1} \middle| \mathbf{g}_t\right). \end{aligned}$$

Therefore, $\mathbb{P}(s_{t+1}|z_t, a_t = 0) = \mathbb{P}(s_{t+1}|s_t, a_t = 0)$. ■

Using Lemma 1, the underlying feedback control problem is an MDP problem based on the $4R$-V state.

Since the state space $\mathcal{S}$ is continuous and thus difficult to work with, we take a common approach to quantize $\mathcal{S}$ into a finite state space[2]. Since the state and action spaces are now finite, the feedback control problem admits an optimal stationary policy that is Markovian and deterministic [25]. Formally, we can represent the optimal stationary policy as $\pi^* = \{d^* : \mathcal{S} \to \mathcal{A}\}^\infty$, where d^* denotes the optimal decision rule and the superscript $^\infty$ indicates that the decision rule d^* is stationary.

B. Challenge of Finding Optimal Policy

Although with the quantized $4R$-V state there exists an optimal policy π^*, finding π^* is challenging because it is difficult to efficiently quantize the state space $\mathcal{S}$. One method is to quantize the $4R$ real spaces individually. Suppose that the quantization level of each real space is L. Then, the total quantization level is L^{4R}, which is large even for moderate values of L and R (*e.g.*, when $L = 4$ and $R = 4$, $L^{4R} \approx 4 \times 10^9$). Thus, we would quickly face the "curse of dimensionality" as L and R increase. Another method is to directly quantize the state space $\mathcal{S}$ by vector quantization technique [30], such as random codebook. For example, a random codebook[3] $\mathcal{C}$ can be designed to include i.i.d. vectors each with the size $4R \times 1$ and the same distribution as s_t. However, to achieve a good performance, the size of $\mathcal{C}$ should be large, which on the other hand would slow down the decision-making procedure.

The above discussion indicates that, perform beamforming feedback control based on the $4R$-V state may be impractical especially when R is large and that the state space reduction is necessary.

IV. State Space Reduction: Analysis on 2-V State

Our focus now is on finding a suitable reduced state for the beamforming feedback control, while trying to maintain the validity of the MDP formulation as much as possible. An attempt of the state space reduction is made in [4]. A 2-V state $s_t = (\|\mathbf{g}_t\|^2, m_t)$ is proposed, where $m_t \triangleq |\mathbf{b}_{u,t-1}^H \mathbf{g}_{u,t}|^2 \in [0, 1]$. The state s_t consists of the current channel power as

[2]For the slow fading channel model, we can approximately treat the quantized version of the problem as an MDP problem.

[3]Note that the codebook here is designed for the state space, and is not for the feedback beamforming vectors.

well as the beamforming gain before the receiver makes the decision. The state space is $\mathcal{S} = \mathcal{R}^+ \times [0,1]$. Based on such a state, upon the receiver's decision, the reward and the state at the next time slot are now as follows:

$$\begin{aligned} a_t = 1 \Rightarrow &\begin{cases} r_t = \log(1 + P\|\mathbf{g}_t\|^2) - c \\ s_{t+1} = (\|\mathbf{g}_{t+1}\|^2, |\mathbf{g}_{u,t}^H \mathbf{g}_{u,t+1}|^2), \end{cases} \\ a_t = 0 \Rightarrow &\begin{cases} r_t = \log(1 + P\|\mathbf{g}_t\|^2 m_t) \\ s_{t+1} = (\|\mathbf{g}_{t+1}\|^2, |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1}|^2) \end{cases} \end{aligned} \tag{10}$$

where q_t is defined in Section III-A. Compared with the $4R$-V state, the size of the 2-V state is largely reduced. However, for the feedback control problem under the 2-V state to be a valid MDP problem, as pointed out in Section III-A, the associated reward function and the transition probabilities should depend on the history only through the current state.

It is claimed in [4] that, the 2-V state is an optimal reduced state[4], and the processes of $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ both form Markov chains and are independent of each other. However, verifying these statements is non-trivial. In this section, we take a detailed look at the validity of the MDP formulation under the 2-V state. In particular, our analysis indicates that the 2-V state is in fact a suboptimal reduction and thus the underlying problem is an approximate MDP. On the other hand, we reveal that the 2-V state provides a good approximation and is efficient.

A. Analysis on $\{\|\mathbf{g}_t\|^2\}$

Based on the channel dynamics in (1), we can show that for finite R, $\{\|\mathbf{g}_t\|^2\}$ is non-Markovian (see Appendix A for a detailed discussion). In this subsection, we demonstrate that, when R is large $\{\|\mathbf{g}_t\|^2\}$ asymptotically forms a Markov chain.

Define $d_t \triangleq \frac{1}{\sqrt{R}}(\|\mathbf{g}_t\|^2 - R)$, which can be treated as a normalized version of $\|\mathbf{g}_t\|^2$. In the following lemma, we give an evolution equation of d_t which will be used later.

Lemma 2: Under the first-order Gauss-Markov channel model (1), $\{d_t\}$ forms a first-order autoregressive process. Formally, we have

$$d_{t+1} = \rho^2 d_t + \tilde{w}_{d,t+1}, \quad t = 0, 1, \cdots$$

where $\tilde{w}_{d,t} \triangleq \frac{1}{\sqrt{R}}\left[2\rho\Re(\mathbf{g}_{t-1}^H \mathbf{w}_t) + \|\mathbf{w}_t\|^2 - (1-\rho^2)R\right]$ and $\{\tilde{w}_{d,t}\} \sim WN(0, 1-\rho^4)$. Furthermore, d_0 is uncorrelated with $\{\tilde{w}_{d,t}\}$, *i.e.,* $\mathbb{E}\left[d_0 \tilde{w}_{d,t}\right] = 0$ for all $t \geq 1$.

Proof: See Appendix B. ∎

Fact 2 below indicates that the Markovian property is transferable for any one-to-one mapping.

Fact 2 ([29]): If $\{x_t\}$ is Markovian and $y_t = g(x_t)$ where $g(\cdot)$ is a one-to-one mapping, then $\{y_t\}$ is also Markovian.

Combining Fact 1, Fact 2, and Lemma 2, we show that $\{\|\mathbf{g}_t\|^2\}$ is asymptotically Markovian in the following proposition.

Proposition 1: Under the first-order Gauss-Markov channel model (1), $\{\|\mathbf{g}_t\|^2\}$ asymptotically forms a Markov chain as $R \to \infty$.

[4]The optimality is in the sense that the reduced 2-V state does not compromise the controller's optimality [4].

Proof: Due to the one-to-one correspondence between $\|\mathbf{g}_t\|^2$ and d_t, showing that $\{\|\mathbf{g}_t\|^2\}$ is Markovian is equivalent to showing that $\{d_t\}$ is Markovian by Fact 2. By Fact 1, to prove that $\{d_t\}$ is Markovian, it suffices to prove that $\{\tilde{w}_{d,t}\}$ is i.i.d. and is independent of d_0. To this end, by Lemma 2, it suffices to show that $(d_0, \{\tilde{w}_{d,t}\})$ is jointly Gaussian. This is because if $(d_0, \{\tilde{w}_{d,t}\})$ were jointly Gaussian, then $\{\tilde{w}_{d,t}\}$ would be an i.i.d. Gaussian sequence because $\{\tilde{w}_{d,t}\} \sim WN(0, 1-\rho^4)$, and d_0 would be independent of $\{\tilde{w}_{d,t}\}$ because $\mathbb{E}\left[d_0 \tilde{w}_{d,t}\right] = 0$ for all $t \geq 1$.

We now proceed to show that $(d_0, \{\tilde{w}_{d,t}\})$ is asymptotically jointly Gaussian as $R \to \infty$ by showing that any linear combination of $d_0, \tilde{w}_{d,t_1}, \tilde{w}_{d,t_2}, \cdots, \tilde{w}_{d,t_n}$ is Gaussian $\forall n > 0$ and $0 < t_1 < t_2 < \cdots < t_n$.

Rewrite

$$\begin{aligned} \tilde{w}_{d,t} &= \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \left[2\rho\Re(\overline{g_{i,t-1}} w_{i,t}) + |w_{i,t}|^2 - (1-\rho^2)\right] \\ &= \frac{1}{\sqrt{R}} \sum_{i=1}^{R} y_{i,t} \end{aligned}$$

where $g_{i,t}$ and $w_{i,t}$ are the i-th elements of $\mathbf{g}_t$ and $\mathbf{w}_t$, respectively, and we have defined $y_{i,t} \triangleq 2\rho\Re(\overline{g_{i,t-1}} w_{i,t}) + |w_{i,t}|^2 - (1-\rho^2)$. Rewrite $d_0 = \frac{1}{\sqrt{R}} \sum_{i=1}^{R} (|g_{i,0}|^2 - 1)$. Let $c_0, c_1, \cdots, c_n$ be the coefficients of the linear combination. Then, we have

$$\begin{aligned} c_0 d_0 + \sum_{l=1}^{n} c_l \tilde{w}_{d,t_l} &= \frac{1}{\sqrt{R}} \left(\sum_{i=1}^{R} c_0 (|g_{i,0}|^2 - 1) + \sum_{l=1}^{n} c_l y_{i,t_l} \right) \\ &= \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \tilde{y}_i \end{aligned} \tag{11}$$

where $\tilde{y}_i \triangleq c_0(|g_{i,0}|^2 - 1) + \sum_{l=1}^{n} c_l y_{i,t_l}$.

By straightforward calculation, we can show that $\{\tilde{y}_i\}$ is i.i.d. with mean zero and variance $c_0^2 + (1-\rho^4)\sum_{i=1}^{n} c_i^2$. Applying the Central Limit Theorem, when $R \to \infty$, the distribution of $\frac{1}{\sqrt{R}} \sum_{i=1}^{R} \tilde{y}_i$ converges to the Gaussian distribution with the same mean and the same variance as those of $\tilde{y}_i$. Hence, $(d_0, \{\tilde{w}_{d,t}\})$ is asymptotically jointly Gaussian.

Therefore, we have $\{\|\mathbf{g}_t\|^2\}$ asymptotically form a Markov chain in the sense that

$$\lim_{R \to \infty} \left| \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2 \right) - \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2 \right) \right| = 0.$$

∎

Using Proposition 1, when R is large, we can approximate $\mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2 \right)$ by $\mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\|\mathbf{g}_t\|^2 \right)$.

B. Analysis on $\{m_t\}$ and Validity of MDP Formulation under 2-V State

In this subsection, we investigate whether the feedback control problem under the 2-V state is an MDP problem.

From (10), the reward function is only a function of the current state, hence satisfying the requirement of an MDP problem. For the transition probabilities, to verify that they depend on the history only through the current state, *i.e.,* (6)

holds, we make the following two assumptions:
A1) The processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ are independent.
A2) R is large.
Under Assumption A1, it is clear that the feedback control problem is not a strict MDP problem since $\{\|\mathbf{g}_t\|^2\}$ is non-Markovian for finite R. This assumption will be discussed later in Section IV-C. Assumption A2 is adopted based on our conclusion in Section IV-A that $\{\|\mathbf{g}_t\|^2\}$ is asymptotically Markovian. Equipped with these two assumptions, we are ready to verify whether (6) holds under the 2-V state.

First assume that $a_t = 1$. Then, $m_{t+1} = |\mathbf{g}_{u,t}^H \mathbf{g}_{u,t+1}|^2$ by (10), and the LHS of (6) is given by

$$\begin{aligned}
&\mathbb{P}(s_{t+1}|z_t, a_t = 1) \\
&= \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2, m_{t+1} \middle| \|\mathbf{g}_0\|^2, m_0, a_0, \cdots, \|\mathbf{g}_t\|^2, m_t, a_t = 1\right) \\
&\approx \mathbb{P}\left(m_{t+1} \middle| m_0, a_0, \cdots, m_t, a_t = 1\right) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right) \qquad (12) \\
&= \mathbb{P}(|\mathbf{g}_{u,t}^H \mathbf{g}_{u,t+1}|^2) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right). \qquad (13)
\end{aligned}$$

Note that (12) is derived by Assumptions A1 and A2 along with Proposition 1. From (12), upon $a_t = 1$, the actual beamforming gain at time slot t reaches the full gain. Hence, the distribution of the successive beamforming gain m_{t+1} is independent of all previous beamforming gains $\{m_\tau\}_{\tau=0}^t$. Therefore, (13) holds.

For the RHS of (6),

$$\begin{aligned}
\mathbb{P}(s_{t+1}|s_t, a_t = 1) &= \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2, m_{t+1} \middle| \|\mathbf{g}_t\|^2, m_t, a_t = 1\right) \\
&= \mathbb{P}(|\mathbf{g}_{u,t}^H \mathbf{g}_{u,t+1}|^2) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right),
\end{aligned}$$

which is equal to (13). As a result, we have (6) approximately hold when $a_t = 1$.

Next assume that $a_t = 0$. Then, $m_{t+1} = |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1}|^2$ by (10), and the LHS of (6) equals

$$\begin{aligned}
&\mathbb{P}(s_{t+1}|z_t, a_t = 0) \\
&\approx \mathbb{P}\left(m_{t+1} \middle| m_0, a_0, \cdots, m_t, a_t = 0\right) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right) \\
&= \mathbb{P}\big(|\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1}|^2 \Big| |\mathbf{g}_{u,-1}^H \mathbf{g}_{u,0}|^2, |\mathbf{g}_{u,1-q_1}^H \mathbf{g}_{u,1}|^2, \\
&\qquad \cdots, |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t}|^2\big) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right) \qquad (14) \\
&= \mathbb{P}\Big(|\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1}|^2 \Big| |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1-q_t}|^2, \\
&\quad |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+2-q_t}|^2, \cdots, |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t}|^2\Big) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right). \qquad (15)
\end{aligned}$$

To see why (15) holds, consider the transition probability of m_{t+1} in (14). By a similar argument as for (13), upon $a_\tau = 1$, there is $m_{\tau+1}$ independent of $\{m_{t'}\}_{t'=0}^\tau$. Hence, for the conditioning part of the transition probability, it suffices to only include those m_τ in which the beamforming vector is the most recent. Thus, (15) is true.

In (15), recall that $q_t \in \{1, \cdots, t+1\}$. To proceed, we consider the following two cases of q_t.

i) $q_t = t + 1$, *i.e.,* no feedback since the beginning. Then the transition probability of m_{t+1} in (15) is

$$\mathbb{P}\Big(|\mathbf{g}_{u,-1}^H \mathbf{g}_{u,t+1}|^2 \Big| |\mathbf{g}_{u,-1}^H \mathbf{g}_{u,0}|^2, |\mathbf{g}_{u,-1}^H \mathbf{g}_{u,1}|^2, \cdots, |\mathbf{g}_{u,-1}^H \mathbf{g}_{u,t}|^2\Big). \qquad (16)$$

ii) $q_t \in \{1, 2, \cdots, t\}$. By the stationarity of the channel gain process, the transition probability of m_{t+1} in (15) equals

$$p\Big(|\mathbf{g}_{u,0}^H \mathbf{g}_{u,q_t+1}|^2 \Big| |\mathbf{g}_{u,0}^H \mathbf{g}_{u,1}|^2, |\mathbf{g}_{u,0}^H \mathbf{g}_{u,2}|^2, \cdots, |\mathbf{g}_{u,0}^H \mathbf{g}_{u,q_t}|^2\Big). \qquad (17)$$

For the RHS of (6),

$$\begin{aligned}
&\mathbb{P}(s_{t+1}|s_t, a_t = 0) \\
&= \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2, m_{t+1} \middle| \|\mathbf{g}_t\|^2, m_t, a_t = 0\right) \\
&= \mathbb{P}\left(|\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t+1}|^2 \middle| |\mathbf{g}_{u,t-q_t}^H \mathbf{g}_{u,t}|^2\right) \mathbb{P}\left(\|\mathbf{g}_{t+1}\|^2 \middle| \|\mathbf{g}_t\|^2\right) \qquad (18)
\end{aligned}$$

where in (18) the value of q_t is unknown because a_{t-1} is unknown.

To make (6) approximately hold when $a_t = 0$, we need to show that (15) and (18) are equal. In other words, we need to demonstrate that the transition probability of m_{t+1} in (18) equals (16) or (17), or equivalently, to demonstrate that $\{m_t\}$ forms a Markov chain when the beamforming vector $\mathbf{b}_{u,t} \equiv \mathbf{g}_{u,-1}$ or $\mathbf{g}_{u,0}$.

Suppose that the beamforming vector $\mathbf{b}_{u,t} \equiv \mathbf{g}_{u,-1}$. By the channel dynamics in (1), the first-order autoregressive model of m_t is given by

$$m_{t+1} = \frac{\rho^2 \|\mathbf{g}_t\|^2}{\|\mathbf{g}_{t+1}\|^2} m_t + \tilde{w}_{m,t+1} \qquad (19)$$

where $\tilde{w}_{m,t+1} \triangleq \frac{2\rho\|\mathbf{g}_t\|}{\|\mathbf{g}_{t+1}\|^2} \Re(\mathbf{g}_{u,t}^H \mathbf{g}_{u,-1} \mathbf{g}_{u,-1}^H \mathbf{w}_{t+1}) + \frac{1}{\|\mathbf{g}_{t+1}\|^2} |\mathbf{g}_{u,-1}^H \mathbf{w}_{t+1}|^2$. Based on [31, Lemma 2] and [31, Lemma 4], it can be shown that the marginal distribution of m_t is given by

$$\mathbb{P}(m_t \le x) = 1 - (1 - x)^{R-1}, \qquad (20)$$

which is non-Gaussian. Since the expression of $\tilde{w}_{m,t+1}$ is complicated and its distribution is non-Gaussian, it is very challenging to rigorously show that $\{m_t\}$ is (asymptotically) Markovian. When the beamforming vector $\mathbf{b}_{u,t} \equiv \mathbf{g}_{u,0}$, similar difficulty arises. Hence, to rigorously establish the Markovian property of $\{m_t\}$ is difficult.

On the other hand, note that when $\mathbf{b}_{u,t} \equiv \mathbf{g}_{u,-1}$ or $\mathbf{g}_{u,0}$, the evolution of $\{m_t\}$ only depends on the evolution of $\{\mathbf{g}_{u,t}\}$. Since $\{\mathbf{g}_t\}$ is Markovian, heuristically, it is reasonable to approximate $\{m_t\}$ as Markovian[5].

Based on the above discussion, under Assumptions A1 and A2, the underlying feedback control problem is an *approximated* MDP problem.

C. Discussion on Dependency between $\{\mathbf{g}_t\}$ and $\{m_t\}$

We now discuss the independence assumption A1. According to the definition, the processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ are independent if the random vectors $\mathbf{g} = [\|\mathbf{g}_{t_1}\|^2, \cdots, \|\mathbf{g}_{t_k}\|^2]$ and $\mathbf{m} = [m_{t'_1}, \cdots, m_{t'_j}]$ are independent for all k, j and all distinct choices of $t_1, \cdots, t_k$ and $t'_1, \cdots, t'_j$, or formally, if

$$\mathbb{P}_{\mathbf{g},\mathbf{m}}(\|\mathbf{g}_{t_1}\|^2, \cdots, \|\mathbf{g}_{t_k}\|^2, m_{t'_1}, \cdots, m_{t'_j})$$

[5]Quantifying the accuracy of such a Markovian model for $\{m_t\}$ is beyond the scope of this paper.

$$= \mathbb{P}_{\mathbf{g}}(\|\mathbf{g}_{t_1}\|^2, \cdots, \|\mathbf{g}_{t_k}\|^2)\mathbb{P}_{\mathbf{m}}(m_{t'_1}, \cdots, m_{t'_j}). \quad (21)$$

Note that the distribution of $\mathbf{m}$ is policy-dependent. For example, if feedback is performed every time slot, all elements in $\mathbf{m}$ are close to 1 with a high probability. In contrast, if feedback is never performed, the marginal distribution of m_t is given by (20). Particularly, from (20), when $R = 2$, m_t is uniformly distributed. Therefore, a strict verification of (21) would require the consideration of all feasible policies, which is obviously intractable.

Instead, we give a heuristic explanation as to why $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ can be treated as independent. Note that these two processes capture two intrinsically distinct aspects of the beamformed system, and one provides little information about the other. Specifically, $\{\|\mathbf{g}_t\|^2\}$ describes the channel power process, with the element $\|\mathbf{g}_t\|^2$ exclusively depending on the current channel; $\{m_t\}$ describes the beamforming gain process and is policy-dependent. For the same channel realization, under different feedback policies, we can have different realizations of $\{m_t\}$ that tell little information about the channel power. On the other hand, from the realization of $\{\|\mathbf{g}_t\|^2\}$, we can barely know the process $\{m_t\}$ either. In the next section, where the feedback control algorithms are proposed, we will revisit the independence assumption A1 and discuss how the feedback algorithms are designed with or without this assumption.

In summary, the above study shows that the 2-V state is not an optimal reduced state because the underlying problem does not rigorously form an MDP problem. Nonetheless, this reduced state is still attractive, since the reduction is efficient and the resultant MDP approximation is justifiable.

D. Quantized 2-V State

Again, since the continuous state space is hard to work with, we need to quantize the 2-V state space. We adopt the quantization method in [4][6]. Specifically, for $\|\mathbf{g}_t\|^2$, we quantize its range $\mathcal{R}^+$ into M levels as $\mathcal{T}_g = \{[0, g'_1), [g'_1, g'_2), \cdots, [g'_{M-1}, +\infty)\}$, where each interval is of probability $1/M$. The representative point of the k-th interval is given by the conditional expectation $\tilde{g}_k = \mathbb{E}\left[\|\mathbf{g}_t\|^2 | \|\mathbf{g}_t\|^2 \in [g'_{k-1}, g'_k)\right]$. For m_t, we quantize its range into N levels as $\mathcal{T}_m = \{[0, 1/N), [1/N, 2/N), \cdots, [(N-1)/N, 1]\}$. The range of m_t is quantized with equal length instead of equal probability because the distribution of m_t is policy-dependent, and thus cannot be determined in advance. The representative point of the i-th interval, denoted by $\tilde{m}_i$, is set to be the mid-point of the associated interval.

For the quantized 2-V state, by treating the underlying feedback control problem as a finite-state MDP problem, there exists an optimal stationary policy that is Markovian and deterministic.

V. Learning Algorithms under the Quantized 2-V State

Based on the quantized 2-V state, we consider practical algorithms for feedback control by treating the underlying problem as an MDP problem. Four learning algorithms, including model-based off-line and on-line algorithms as well as a model-free on-line algorithm, are provided.

A. Model-Based Off-Line PI Algorithms

In model-based learning, the receiver first derives the transition probabilities and then obtains an optimal feedback policy employing policy iteration (PI) [25].

1) Learning of Transition Probabilities: We now discuss how to estimate transition probabilities. Denote the state transition probability as $\mathbb{P}(s'|s, a)$, where the time indexes are suppressed for ease of notation. When learning the transition probabilities, we consider the processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ to be either independent or correlated. In particular, we call the algorithm under the independence assumption the "independent" off-line PI algorithm, and call the other the "joint" off-line PI algorithm.

Denote the sampling transition probabilities of $\|\mathbf{g}_t\|^2$, m_t, and $(\|\mathbf{g}_t\|^2, m_t)$ by $\mathbb{P}_g(\cdot), \mathbb{P}_m(\cdot)$, and $\mathbb{P}_{gm}(\cdot)$, respectively, which can be obtained by the maximum likelihood estimation. For the "independent" off-line PI algorithm, $\mathbb{P}(s'|s, a)$ is derived by the product of $\mathbb{P}_g(\cdot)$ and $\mathbb{P}_m(\cdot)$; for the "joint" off-line PI algorithm, $\mathbb{P}(s'|s, a)$ is derived based on $\mathbb{P}_{gm}(\cdot)$. Note that for slow fading environment, we can assume that upon feedback, m_{t+1} will achieve the highest beamforming gain, *i.e.,* jump to the highest quantized state N. Under this assumption, we summarize the procedure of generating the transition probabilities for both the "joint" and "independent" off-line PI algorithms in Appendix C.

Since learning transition probabilities incurs data rate loss, it is desirable to make the learning phase as short as possible. In the "independent" off-line PI algorithm, we simultaneously learn two transition probability matrices $\mathbb{P}_g(\cdot)$ and $\mathbb{P}_m(\cdot)$, whose sizes are $M \times M$ and $N \times N$, respectively. In the "joint" off-line PI algorithm, we learn the transition probability matrix $\mathbb{P}_{gm}(\cdot)$, whose size is $MN \times MN$. This size is much larger than that of $\mathbb{P}_g(\cdot)$ or $\mathbb{P}_m(\cdot)$. Therefore, a longer learning phase is required for the "joint" algorithm to achieve a similar learning accuracy as that of the "independent" one. In this sense, the "joint" algorithm is less efficient.

2) PI: Given the transition probabilities, we can form the expected total discounted reward (value function) under the state s and policy π as

$$V^{\pi}(s) = r(s, a) + \lambda \sum_{s'} \mathbb{P}\left(s' \middle| s, a\right) V^{\pi}(s') \quad (22)$$

where the action in state s follows the policy π, and the time indexes are suppressed for ease of notation. Denote π^* as an optimal policy. Then, the optimal value function should satisfy the Bellman's equation [25]

$$V^{\pi^*}(s) = \max_{a \in \mathcal{A}} \left[r(s, a) + \lambda \sum_{s'} \mathbb{P}\left(s' \middle| s, a\right) V^{\pi^*}(s') \right]. \quad (23)$$

Using PI to obtain a policy of (23) involves alternating between policy evaluation and policy improvement iteratively until the algorithm converges [25]. For finite state and action spaces, it has been proven that PI is guaranteed to converge to

[6]Note that, to get better performance the joint quantization scheme needs to be considered. But this is beyond the scope of this paper.

an optimal solution within finite iterations. In our simulation, with the quantization levels $M = N = 4$, typically 2 or 3 iterations are needed for convergence, while with $M = N = 32$, typically 7 or 8 iterations are needed. For PI, the computational complexity mainly lies in policy evaluation and is of the order $O((MN)^3)$. The required storage is largely determined by the size of the transition probability matrix and is of the order $O((MN)^2)$.

B. Model-Free On-Line EQ Algorithm

As opposed to model-based learning, in model-free learning, the receiver tries to learn an optimal policy π^* directly without deriving $\mathbb{P}(s'|s,a)$ first. In this subsection, we adapt a model-free algorithm, the Exploratory Q-learning (EQ) algorithm, for feedback control.

1) Background of Q-Learning: Q-learning is a model-free algorithm, by which the decisions are made through the learning of the optimal Q-factors [32]. Denote $Q^\pi(s,a)$ as the Q-factor of the state-action pair (s,a) under a policy π. It is given by

$$Q^\pi(s,a) = r(s,a) + \lambda \sum_{s'} \mathbb{P}(s'|s,a) V^\pi(s') \tag{24}$$

where $V^\pi(s')$ is the value function of the state s' under the policy π. From (24), $Q^\pi(s,a)$ represents the expected total discounted reward by taking the action a in the state s and following the fixed policy π thereafter.

If π^* is an optimal policy, we have $V^{\pi*}(s) = \max_{a\in\mathcal{A}} Q^{\pi*}(s,a)$, which means that π^* can be derived from the optimal Q-factors $Q^{\pi*}(s,a)$. To learn the optimal Q-factors, at each decision epoch t, the decision maker first observes the current state s, then selects an action and observes the next state s'. Denote $Q_t(s,a)$ as the Q-factor at decision epoch t. The update of $Q_t(s,a)$ is given by

$$Q_t(s,a) = \begin{cases} (1-\alpha_t(s,a))Q_{t-1}(s,a) + \alpha_t(s,a)\left[r_t(s,a) + \lambda V_{t-1}(s')\right], & \text{if } s = s_t \text{ and } a = a_t \\ Q_{t-1}(s,a), & \text{otherwise} \end{cases} \tag{25}$$

where $V_{t-1}(s') = \max_{a\in\mathcal{A}} Q_{t-1}(s',a)$ and $\alpha_t(s,a)$ is the learning rate of the state-action pair (s,a) at time slot t. Note that, by (25), only the Q-factor associated with the *visited* state-action pair is updated. It is proven in [32] that, if all state-action pairs are visited infinitely often and the learning rate is appropriately designed, all Q-factors will eventually converge to the optimal ones.

2) EQ Algorithm: The EQ algorithm is introduced in [33]. It combines Q-learning with a counter-based directed exploration strategy. Below, we briefly summarize how the EQ algorithm chooses an action at each decision epoch.

Denote $n_t(s,a)$ as the number of times that the state-action pair (s,a) is visited up to time slot t, and $n_t(s)$ as the number of times that the state s is visited up to time slot t. Let $c_t(s,a)$ reflect the number of times that the action a has *not* been performed in the state s since the initial time. Initialize $c_0(s,a) = 0$, $\forall s, a$.

In the EQ algorithm, given the current state s_t, a *greedy action* $\hat{a}_t = \arg\max_{a\in\mathcal{A}} Q_t(s_t,a)$ is first determined, which leads to the maximum Q-factor at time slot t. The final action a_t is decided by the following maximization problem:

$$\max_{a\in\mathcal{A}} \quad g_t(s_t,a) \triangleq \begin{cases} \frac{c_t(s_t,a)}{n_t(s_t,a)} + 1, & \text{if } a = \hat{a}_t \\ \frac{c_t(s_t,a)}{n_t(s_t,a)}, & \text{otherwise} \end{cases} \tag{26}$$

where the addition of 1 indicates the decision maker's preference for the greedy action. It is possible that the decision maker finally takes a *non-greedy action* $a \neq \hat{a}_t$ at time slot t. This could happen provided that the ratio $\frac{c_t(s_t,a)}{n_t(s_t,a)}$ is large, *i.e.,* the action a has not been performed in s_t for a long time. The update of $c_t(s_t,a)$ is as follows:

$$c_{t+1}(s_t,a) = \begin{cases} c_t(s_t,a) + \Theta(n_{t+1}(s_t) - n_{t+1}(s_t,a)), & \text{if } a \neq a_t \\ c_t(s_t,a), & \text{otherwise} \end{cases} \tag{27}$$

where $\Theta(n)$ is a positive function satisfying the conditions

$$\lim_{n\to\infty} \Theta(n) = 0, \text{ and } \sum_{n=1}^{\infty} \Theta(n) = \infty. \tag{28}$$

a) Complexity: At each decision epoch, given the current state, the operation of the EQ algorithm is based on $|\mathcal{A}|$ state-action pairs. Thus, the computational complexity is of the order $O(|\mathcal{A}|)$. In our problem, since there are only two possible actions, the computational complexity of the EQ algorithm is low. The state-action based parameters, such as the Q-factors, should be stored for each update. Thus the required storage size of the EQ algorithm is of the order $O(|\mathcal{A}|MN)$.

b) Design parameters: For the updates in (25) and (27), two parameters need to be carefully designed: the learning rate $\alpha_t(s,a)$ and the positive function $\Theta(n)$. In [27], the learning rate is suggested as $\alpha_t(s,a) = \frac{\alpha_0 \tau}{\tau + n_t(s,a)}$, where α_0 is the initial learning rate and τ is some positive parameter. The positive function $\Theta(n)$ is set to be $\frac{1}{n^{1/\theta}}$, $\theta \geq 1$. From (27), a larger θ can result in a larger value of $c_{t+1}(s_t,a)$ for $a \neq a_t$ (usually a non-greedy action). Therefore, the value of θ affects the trade-off between exploration (a non-greedy action) and exploitation (the greedy action). In Section VI, we will study how to choose θ and α_0.

C. Model-Based On-Line PI Algorithm

For comparison, we also provide a model-based on-line PI algorithm, which combines PI with a counter-based directed exploration strategy similar to that in the EQ algorithm [33]. We briefly describe this algorithm below.

In this algorithm, at each decision epoch, first, the estimates of the transition probabilities $\mathbb{P}(s'|s,a)$ are updated by the observation history $\{\|\mathbf{g}_\tau\|^2, m_\tau, a_\tau\}_{\tau=0}^t$ where the processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ are jointly considered; second, with the updated $\mathbb{P}(s'|s,a)$, PI is employed to generate the greedy action $\hat{a}_t$; and third, the final action a_t is determined by the optimization problem (26).

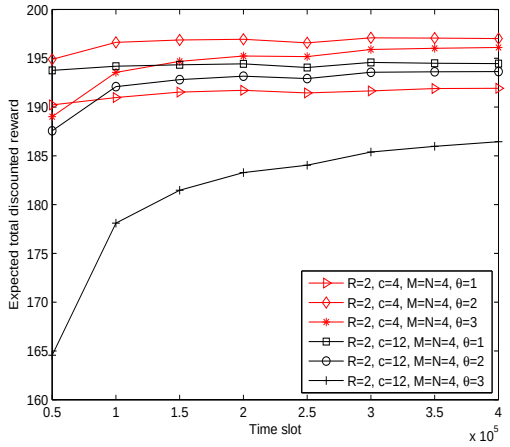


Fig. 1. Design of θ: 2-V EQ algorithm with $R = 2$, $M = N = 4$, $c = 4$ and 12, and $\theta = 1, 2$ and 3.

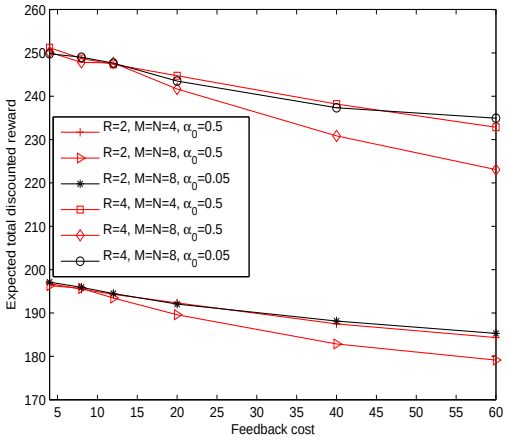


Fig. 2. Design of α_0: 2-V EQ algorithm with $R = 2$ and 4, $M = N = 4$ and 8, and $\alpha_0 = 0.5$ and 0.05.

Compared with the EQ algorithm in which the greedy action $\hat{a}_t$ is generated by Q-learning, in the model-based on-line PI algorithm $\hat{a}_t$ is produced by PI. Compared with the model-based "joint" off-line PI algorithm which generates the transition probabilities once, in the on-line PI algorithm, the estimates of $\mathbb{P}(s'|s, a)$ are updated at every decision epoch by incorporating newly obtained observation data. Thus, in using these data, the algorithm seeks a certain balance between exploration and exploitation. Furthermore, the computational complexity of the on-line PI algorithm is high, since PI is performed at every decision epoch.

VI. Performance Study of Proposed Learning Algorithms

In this section, we will study and compare all learning algorithms proposed in the previous section.

In all simulation, we set the discount factor $\lambda = 0.98$, the normalized Doppler frequency $f_D T_s = 0.01$, and the power at the transmitter $P = 10$ dBW (except otherwise mentioned). The expected total discounted reward in (4) is approximated by the sample mean of the total discounted rewards over 400 independently generated channel sequences. For the n-th generated channel sequence, the total discounted reward beginning at time slot t is approximated by $\hat{V}_t^{(n)} = \sum_{\tau=t}^{t+\Delta t} \lambda^{\tau-t} r_\tau^{(n)}$, where $\Delta t = 350$. The parameter τ in the learning rate $\alpha_t(s, a)$ of the EQ algorithm is set to be 300. The feedback cost is drawn from the set $\{4, 8, 12, 20, 40, 60\}$. Note that these values are simply chosen as examples in our simulation.

A. Model-Based vs. Model-Free Algorithms

We first provide the guideline on choosing the parameters θ and α_0 in the EQ algorithm. Then we compare the EQ algorithm and the on-line PI algorithm.

1) EQ Algorithm – Design of θ and α_0: Let $R = 2$ and $M = N = 4$. The feedback cost is set to be $c = 4$ and 12. In Fig. 1, we show the expected total discounted reward under the EQ algorithm with $\theta = 1, 2$, and 3 (recall that θ controls the exploration-exploitation trade-off in the EQ algorithm). Each curve in Fig. 1 displays the convergence behavior associated with a specified θ. We see that, when the feedback cost is low and equals 4, $\theta = 2$ (a balanced exploration-exploitation) results in the best system performance, while $\theta = 1$ (too much exploitation) results in the lowest performance. When the feedback cost is high and equals 12, $\theta = 1$ results in the best performance, and larger values of θ (more exploration) prolong the convergence time of the EQ algorithm. In summary, the choice of θ depends on the feedback cost. When the feedback cost is low, the system benefits more from exploration of the non-greedy action, and thus a larger θ is preferred. When the feedback cost is high, such exploration becomes costly, and thus a smaller θ is preferred.

In Fig. 2, for $R = 2$ and 4, we investigate the effect of quantization levels on the design of the initial learning rate α_0. We display the expected total discounted reward under two sets of quantization levels and various feedback costs. The parameter θ is set appropriately. When $M = N = 4$, $\alpha_0 = 0.5$ is found to produce the best performance. However, when $M = N = 8$, $\alpha_0 = 0.5$ is shown to be inferior to $\alpha_0 = 0.05$, especially at the high feedback cost region. For example, when $c = 60$, the degradation of $\alpha_0 = 0.5$ regarding $\alpha_0 = 0.05$ is $3\%(R = 2)$ and $5\%(R = 4)$. Hence, Fig. 2 indicates that, the initial learning rate α_0 should be adjusted to the quantization levels. Specifically, α_0 should be smaller when the quantization levels are high. This is because higher quantization levels result in less observation samples for each quantized level, and consequently, the learning process needs to slow down. Furthermore, we observe that, if α_0 is appropriately designed, the improvement by increasing quantization levels is limited.

2) EQ Algorithm vs. On-Line PI Algorithm: In Fig. 3, for $R = 2$ and 4, and $M = N = 4$, we compare the EQ algorithm and the model-based on-line PI algorithm under various feedback costs. Recall that the only difference between these two algorithms is how to generate the greedy action. We see that, the EQ algorithm outperforms the on-line PI algorithm at the mediate-to-high cost region, showing an advantage of the EQ algorithm. The advantage of the EQ algorithm could be attributed to the reward-driven nature of

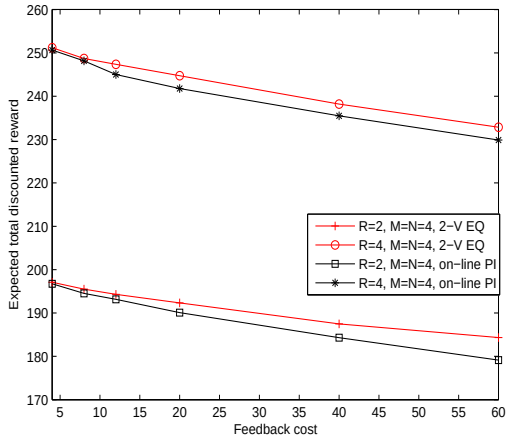


Fig. 3. EQ algorithm vs. on-line PI algorithm: $R = 2$ and 4, $M = N = 4$.

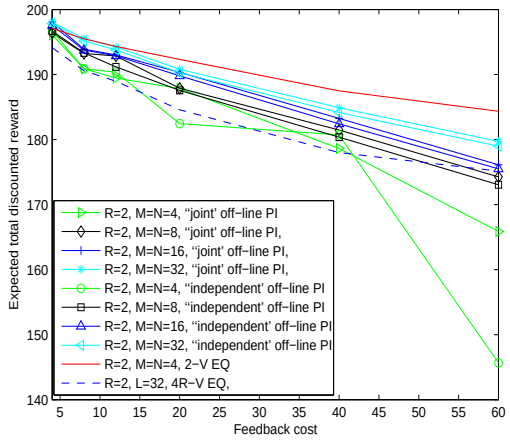


Fig. 4. "Joint" off-line PI algorithm vs. "independent" off-line PI algorithm: $R = 2$, $P = 10$ dBW, and $M = N = 4, 8, 16$, and 32.

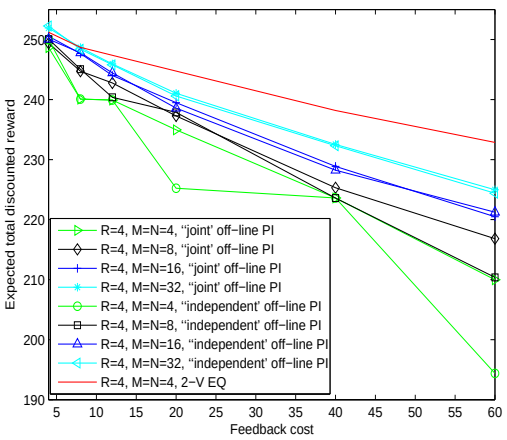


Fig. 5. "Joint" off-line PI algorithm vs. "independent" off-line PI algorithm: $R = 4$, $P = 10$ dBW, and $M = N = 4, 8, 16$ and 32.

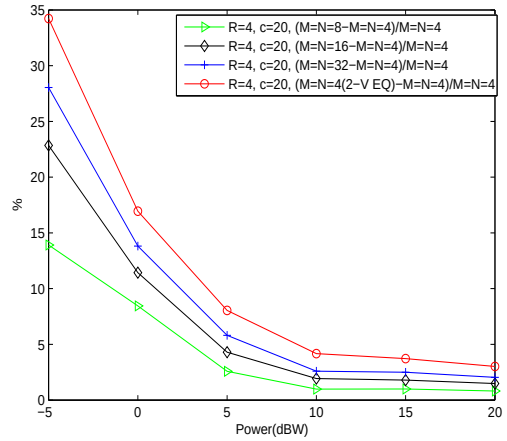


Fig. 6. "Joint" off-line PI algorithm: $R = 4$, $c = 20$, and $M = N = 4$.

Q-learning.

B. "Independent" vs. "Joint" Off-Line PI Algorithms

In Section IV we have argued that determining whether the processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ are independent is challenging. In the proposed two off-line PI algorithms, these two processes are treated either jointly or independently. In this subsection, we compare these two algorithms under two scenarios: a fixed transmit power level with various feedback costs, and various transmit power levels with a fixed feedback cost.

1) Effect of Feedback Cost: In Fig. 4, with the transmit power being $P = 10$ dBW, we compare the expected total discounted reward of the "joint" and "independent" off-line PI algorithms under various feedback costs. We set $R = 2$ and the quantization levels $M = N = 4, 8, 16$, and 32. We observe that, the "joint" algorithm outperforms the "independent" one in general, which indicates that $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ are correlated. Furthermore, as the quantization levels increase, performance of both algorithms improves, and also the performance gap between these two reduces. In particular, when $M = N = 32$, the performance gap becomes negligible, indicating that treating $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ as independent is acceptable. For comparison, the EQ algorithm with $M = N = 4$ is shown. Surprisingly, it yields the best performance compared with all off-line PI algorithms, especially at the high feedback cost region.

In Fig. 4, we additionally include the EQ algorithm designed for the original $4R$-V state. The average quantization level per real variable is set to be $L = 32$, and a random code book with the size 256 is used. Compared with the algorithms designed for the 2-V state, the $4R$-V state based EQ algorithm is inferior to the "independent" off-line PI algorithm with $M = N = 16$. This observation reveals that, although the $4R$-V state is optimal, there is a large performance loss associated with quantization. Therefore, the 2-V state is more efficient.

The same experiments are conducted for $R = 4$ in Fig. 5. From this figure, we can see similar observations as those in Fig. 4. In particular, the performance gap between the "joint" and "independent" off-line PI algorithms is small for $M = N = 16$ and 32. This again indicates that assuming $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ to be independent is reasonable.

2) Effect of Transmit Power: In Fig. 6, with a fixed feedback cost $c = 20$, we show performance of the model-based "joint" off-line PI algorithm under various transmit power levels. Using the quantization levels $M = N = 4$

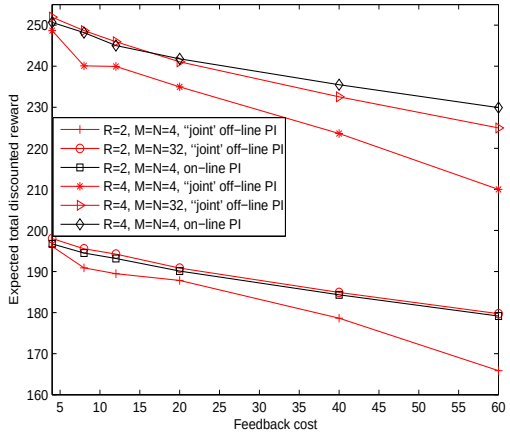


Fig. 7. Off-line PI algorithm vs. on-line PI algorithm: $R = 2$ and 4, and $M = N = 4$ and 32.

as the base, we plot the performance improvement under $M = N = 8, 16$, and 32. We see that, performance improves with the quantization levels, and the improvement is significant at the low transmit power region. For example, when $P = -5$ dBW, the advantages of $M = N = 8, 16$, and 32 over the base are as high as $14\%, 23\%$, and 28%, respectively. In contrast, when $P = 20$ dBW, the advantages are only $0.8\%, 1.5\%$, and 2%, respectively. The reason for such a significant improvement at the low transmit power region (corresponding to the low average SNR) is as follows. At the lower power region, which is the linear region of the capacity, the increased SNR directly translates to the capacity gain thus the throughput gain. At the high power region, which is the non-linear region of the capacity, the increased SNR translates to a smaller throughput improvement. For comparison, we also show performance of the EQ algorithm, which is observed to have the highest improvement regarding the base. The observations of the "independent" off-line PI algorithm are similar and are omitted here.

C. Off-Line PI vs. On-Line PI Algorithms

In Fig. 7, for $R = 2$ and 4, and $M = N = 4$ and 32, we compare the "joint" off-line PI algorithm with the on-line PI algorithm under various feedback costs. We see that, with low quantization levels, *e.g.*, $M = N = 4$, the off-line PI algorithm is inferior to the on-line PI algorithm. Specifically, when $c = 60$, the degradation percentages regarding the on-line algorithm are 7% $(R = 2)$ and 9% $(R = 4)$. In contrast, with high quantization levels, *e.g.*, $M = N = 32$, performance of the off-line algorithm improves and becomes comparable with respect to that of the on-line algorithm under $M = N = 4$. This indicates that the on-line PI algorithm can tolerate a higher quantization error than the off-line PI algorithm.

D. Discussions

In Table I, we summarize the time efficiency, the (storage) space efficiency, and the computational efficiency of all proposed algorithms.

For the "joint" and "independent" off-line PI algorithms, the time efficiency is considered based on the number of the time slots spent on generating transition probabilities. As discussed in Section V-A1, the "independent" algorithm is more efficient than the "joint" one. For the on-line PI algorithm and the EQ algorithm, since there is no extra learning phase incurred, the time efficiency is considered based on the number of the time slots required for achieving a stable performance. Simulation shows that the latter two algorithms consume less number of time slots than the former two algorithms and thus are more efficient.

In Section VI-A, we compared two on-line algorithms and observed that the EQ algorithm has an advantage over the on-line PI algorithm. Also, from Table I, the EQ algorithm has a higher space efficiency and computational efficiency than the on-line PI algorithm. Therefore, as far as on-line algorithms are concerned, the EQ algorithm is preferable.

In Section VI-B, we compared the "joint" and "independent" off-line PI algorithms under various transmit power levels and feedback costs. We observed that these two algorithms have a similar performance under high quantization levels. This observation indicates that the processes $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$ can be approximately treated as independent. Therefore, when off-line policy is considered, the "independent" off-line algorithm is preferable due to its higher time efficiency.

In Section VI-C, we compared two PI-based algorithms (on-line/off-line) and observed that the on-line PI algorithm is superior to the "joint" off-line PI algorithm. This observation shows that, considering the trade-off between exploration and exploitation, the transition probabilities derived by leveraging on-line data are more accurate.

To summarize, we suggest the following rule for choosing an appropriate algorithm among all proposed ones. If the channel is statistically stable, *i.e.,* the statistics of $\mathbf{g}_0$ or $\mathbf{w}_t$ does not change, the "independent" off-line PI algorithm is preferable. When used, the "independent" algorithm should adopt high quantization levels especially at high feedback cost region and low transmit power region. Also, to avoid repeated generation, the derived policy can be stored for future use. In contrast, if the channel is statistically unstable, or if the time/space/computational efficiencies are the main concerns, the EQ algorithm is the best candidate.

VII. Further Discussion on Stochastic Feedback: Practical Implementation and Performance

In this work, we focus on the timing of feedback while assuming the feedback beamforming vectors to be perfect. In practice, feedback beamforming vectors need to be quantized using a codebook. The stochastic feedback control considered in this work can be combined with different codebook-based strategies, such as traditional codebook quantization [6], [7], the compression method [18]–[20], and differential vector quantization [13], [14], [21], [22], for practical feedback implementation of beamforming vectors. The value of the feedback cost c in (3) may depend on the size of the codebook. Since the codebook design is not the focus of this work, the determination of c is beyond the scope of this paper, and the

TABLE I
EFFICIENCY ANALYSIS OF ALL PROPOSED ALGORITHMS

	"Joint" off-line PI	"Independent" off-line PI	On-line PI	EQ
Time efficiency	Low	Medium	High	High
Space efficiency	Low ($O((MN)^2)$)	Low ($O((MN)^2)$)	Low ($O((MN)^2)$)	High ($O(\|\mathcal{A}\|MN)$)
Computational efficiency (per run)	Low ($O((MN)^3)$)	Low ($O((MN)^3)$)	Low ($O((MN)^3)$)	High ($O(\|\mathcal{A}\|)$)

design of codebook-based stochastic feedback remains open for future study.

The stochastic feedback control results in aperiodic feedback which needs appropriate signaling and reporting designs for practical implementation. The current practical systems such as LTE [34] have built-in aperiodic feedback reporting modes, besides periodic ones. They can be used to provide need-based reporting of CSI and/or beamforming vector index for multi-antenna transmission whenever the channel condition changes. Such aperiodic feedback in LTE is carried through the uplink data channel. However, the signaling is initiated by base stations, not users, while our stochastic control requires user-triggered signaling. Some novel design can be made to accomplish the signaling of the stochastic feedback scheme. For example, we could design a 1-bit on-off signaling in the uplink control channel with "on" signal indicating the instance of feedback[7].

Regarding the system performance, comparison between stochastic feedback control and periodic feedback control has been studied in [4]. Under a similar system model as ours, the authors compare the stochastic feedback scheme with a periodic feedback scheme of which the period is optimized. Through extensive simulation, the stochastic feedback has been shown to outperform the periodic one.

Furthermore, in this work, our design objective is maximizing the overall throughput, which directly factors in the rate loss due to feedback. In contrast, traditional codebook-based periodic feedback is a type of static scheme. The design objective does not involve feedback cost, and thus does not lead to optimizing the overall throughput. As such, stochastic feedback control will outperform the periodic feedback scheme (assuming the same quantization and codebook used), due to its design goal of overall throughput maximization.

VIII. CONCLUSION

Based on a first-order Gauss-Markov channel model, we have considered the stochastic feedback control of beamforming vector for MISO systems, and provided practical algorithms for implementation. We have showed that, although based on a $4R$-V state the underlying feedback control problem can be formulated as an MDP problem, finding an optimal policy is challenging. We have then considered a reduced 2-V state and studied the validity of its MDP formulation. Our investigation indicates that the reduced 2-V state is in fact suboptimal but meanwhile is justifiable and efficient. Based on the quantized 2-V state, we have provided and analyzed four learning algorithms. Through a detailed study of all proposed algorithms, we have suggested an application rule for choosing an appropriate algorithm under different channel conditions and efficiency concerns.

[7]Such signaling channel can be designed similarly as the ACK/NACK channel in LTE.

APPENDIX A
NON-MARKOVIAN PROPERTY OF $\{\|\mathbf{g}_t\|^2\}$ FOR FINITE R

From the definition, showing that $\{||\mathbf{g}_t||^2\}$ (or $\{||\mathbf{g}_t||\}$) forms a Markov chain is equivalent to showing that

$$\mathbb{P}(\|\mathbf{g}_{t+1}\|^2 \Big| \ \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2) = \mathbb{P}(\|\mathbf{g}_{t+1}\|^2 \Big| \ \ \|\mathbf{g}_t\|^2). \quad (29)$$

Based on $\|\mathbf{g}_{t+1}\|^2 = \rho^2\|\mathbf{g}_t\|^2 + \|\mathbf{w}_{t+1}\|^2 + 2\rho\|\mathbf{g}_t\|\Re(\mathbf{g}_{u,t}^H\mathbf{w}_{t+1})$ and $\mathbf{g}_t = \|\mathbf{g}_t\|\mathbf{g}_{u,t}$, showing (29) is equivalent to showing

$$\begin{aligned}&\mathbb{P}\Big(\rho^2\|\mathbf{g}_t\|^2 + \|\mathbf{w}_{t+1}\|^2 + 2\rho\|\mathbf{g}_t\|\Re(\mathbf{g}_{u,t}^H\mathbf{w}_{t+1})\Big|\|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2\Big)\\ &= \mathbb{P}\left(\rho^2\|\mathbf{g}_t\|^2 + \|\mathbf{w}_{t+1}\|^2 + 2\rho\|\mathbf{g}_t\|\Re(\mathbf{g}_{u,t}^H\mathbf{w}_{t+1}) \ \Big|\|\mathbf{g}_t\|^2\right).\end{aligned} \quad (30)$$

Note that the conditional probability of $\mathbf{g}_{u,t}^H\mathbf{w}_{t+1}$ in (30) is given by

$$\begin{aligned}&\mathbb{P}\left(\mathbf{g}_{u,t}^H\mathbf{w}_{t+1}\Big| \ \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2\right)\\ &= \int \mathbb{P}(\mathbf{g}_{u,t}\Big| \ \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2)\mathbb{P}(\mathbf{g}_{u,t}^H\mathbf{w}_{t+1}\Big| \ \mathbf{g}_{u,t}) \ d\mathbf{g}_{u,t}.\end{aligned} \quad (31)$$

Recall that $\mathbf{g}_{u,t} = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}$. Now we are left to check whether $\mathbb{P}(\frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}\Big| \ \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2) = \mathbb{P}(\frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}\Big| \ \|\mathbf{g}_t\|^2)$, or equivalently, whether

$$\mathbb{P}(\mathbf{g}_t\Big| \ \|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_t\|^2) = \mathbb{P}(\mathbf{g}_t\Big| \ \|\mathbf{g}_t\|^2). \quad (32)$$

Note that $\mathbf{g}_t$ is a function of $\|\mathbf{g}_{t-1}\|$ through $\mathbf{g}_t = \rho\|\mathbf{g}_{t-1}\|\mathbf{g}_{u,t-1} + \mathbf{w}_t$; while $\|\mathbf{g}_t\|^2$ is related to $\|\mathbf{g}_{t-1}\|$ through a different function that $\|\mathbf{g}_t\|^2 = \rho^2\|\mathbf{g}_{t-1}\|^2 + \|\mathbf{w}_t\|^2 + 2\rho\Re\{\mathbf{g}_{t-1}^H\mathbf{w}_t\}$. Thus, intuitively, $\|\mathbf{g}_t\|^2$ does not contain all information needed to determine $\mathbf{g}_t$ as compared to the condition $(\|\mathbf{g}_0\|^2, \cdots, \|\mathbf{g}_{t-1}\|^2, \|\mathbf{g}_t\|^2)$ given in the LHS of (32). Therefore, $\{\|\mathbf{g}_t\|^2\}$ is not Markovian for finite R.

APPENDIX B
PROOF OF LEMMA 2

Based on (1), we have

$$\begin{aligned}\|\mathbf{g}_{t+1}\|^2 &= \|\rho\mathbf{g}_t + \mathbf{w}_{t+1}\|^2\\ &= \rho^2\|\mathbf{g}_t\|^2 + \rho\mathbf{g}_t^H\mathbf{w}_{t+1} + \rho\mathbf{w}_{t+1}^H\mathbf{g}_t + \|\mathbf{w}_{t+1}\|^2.\end{aligned} \quad (33)$$

By the definitions of d_t and $\tilde{w}_{d,t}$, from (33), we have

$$d_{t+1} = \rho^2 d_t + \tilde{w}_{d,t+1}.$$

To show $\{\tilde{w}_{d,t}\} \sim WN(0, 1-\rho^4)$, we need to show that, in $\{\tilde{w}_{d,t}\}$, all distinct elements are uncorrelated and all elements are with mean zero and variance $1-\rho^4$. From (1), $\mathbf{w}_t$ is independent of $\mathbf{g}_{t'}$ whenever $t > t'$. Using this fact, there is $\mathbb{E}\left[\tilde{w}_{d,t}\right] = 0$. Define

$$\begin{aligned} w'_{d,t} &\triangleq \tilde{w}_{d,t} + \sqrt{R}(1-\rho^2) \\ &= \frac{1}{\sqrt{R}}\left(\rho \mathbf{g}_{t-1}^H \mathbf{w}_t + \rho \mathbf{w}_t^H \mathbf{g}_{t-1} + \|\mathbf{w}_t\|^2\right). \end{aligned}$$

It can be calculated that $\mathbb{E}\left[w'_{d,t}\right] = \sqrt{R}\sigma^2$, $\mathbb{E}\left[w'^2_{d,t}\right] = 2\rho^2\sigma^2 + (R+1)\sigma^4$, and $\mathbb{E}\left[w'_{d,t}w'_{d,t+q}\right] = R\sigma^4$ for $q \neq 0$. Using the one-to-one correspondence between $w'_{d,t}$ and $\tilde{w}_{d,t}$, we have $\mathrm{var}(\tilde{w}_{d,t}) = \mathbb{E}\left[\tilde{w}^2_{d,t}\right] = 1-\rho^4$, and $\mathbb{E}\left[\tilde{w}_{d,t}\tilde{w}_{d,t+q}\right] = 0$ for $q \neq 0$. Therefore, $\tilde{w}_{d,t} \sim WN(0, 1-\rho^4)$. Also, by straightforward calculation,

$$\begin{aligned} \mathbb{E}\left[d_0 \tilde{w}_{d,t}\right] = \frac{1}{R}\mathbb{E}\Big[&(\|\mathbf{g}_0\|^2 - R)\big[\rho \mathbf{g}_{t-1}^H \mathbf{w}_t \\ &+ \rho \mathbf{w}_t^H \mathbf{g}_{t-1} + \|\mathbf{w}_t\|^2 - (1-\rho^2)R\big]\Big] = 0 \end{aligned} \quad (34)$$

for all $t \geq 1$. Since d_0 and all elements of $\{\tilde{w}_{d,t}\}$ are with mean zero, from (34), d_0 and $\{\tilde{w}_{d,t}\}$ are uncorrelated.

Appendix C
Generation of Transition Probabilities for "Joint" and "Independent" Off-Line PI Algorithms

Under the quantized 2-V state space, in Algorithm 1 we summarize the procedure of generating the sampling transition probabilities of $\|\mathbf{g}_t\|^2$, m_t, and $(\|\mathbf{g}_t\|^2, m_t)$ without feedback. We denote $s = (k,i)$ and $s' = (h,j)$ as shorthands of $s = (\tilde{g}_k, \tilde{m}_i)$ and $s' = (\tilde{g}_h, \tilde{m}_j)$, respectively, and denote $\mathcal{T}_g(k)$ as the k-th set of $\mathcal{T}_g$ and $\mathcal{T}_m(i)$ as the i-th set of $\mathcal{T}_m$.

Using $\mathbb{P}_g$ in (37) and $\mathbb{P}_m$ in (38) we derive the transition probabilities for the "independent" off-line PI algorithm as

$$\mathbb{P}\left((h,j)\middle|(k,i),a\right) = \begin{cases} \mathbb{P}_g(h|k)\mathbb{P}_m(j|i), & a=0 \\ \mathbf{1}(j=N)\mathbb{P}_g(h|k), & a=1 \end{cases}. \quad (35)$$

Using $\mathbb{P}_{gm}$ in (39) we derive the transition probabilities for the "joint" off-line PI algorithm as

$$\begin{aligned} &\mathbb{P}\left((h,j)\middle|(k,i),a\right) \\ &= \begin{cases} \mathbb{P}_{gm}((h,j)|(k,i)), & a=0 \\ \mathbf{1}(j=N)\left[\sum_{q=1}^{N}\mathbb{P}_{gm}((h,q)|(k,i))\right], & a=1 \end{cases}. \end{aligned} \quad (36)$$

Algorithm 1 Generation of sampling transition probabilities of $\|\mathbf{g}_t\|^2$, m_t, and $(\|\mathbf{g}_t\|^2, m_t)$ without feedback.

1: Initialization: $n=1$, $\mathbb{P}_g = \mathbf{0}_{M,M}$, $\mathbb{P}_m = \mathbf{0}_{N,N}$, and $\mathbb{P}_{gm} = \mathbf{0}_{MN,MN}$.

2: a) Based on (1), generate a sequence of $\{\mathbf{g}_t\}$ of length T_c. Letting $\mathbf{b}_{u,t} \equiv \mathbf{g}_{u,0}$, obtain the sequences of $\{\|\mathbf{g}_t\|^2\}$ and $\{m_t\}$.

b) Derive the sampling transition probabilities:

$$\begin{aligned} &\mathbb{P}_g^{(n)}(h|k) \\ &= \frac{\sum_{t=0}^{T_c-2}\mathbf{1}\left(\|\mathbf{g}_t\|^2 \in \mathcal{T}_g(k), \|\mathbf{g}_{t+1}\|^2 \in \mathcal{T}_g(h)\right)}{\sum_{t=0}^{T_c-1}\mathbf{1}\left(\|\mathbf{g}_t\|^2 \in \mathcal{T}_g(k)\right)}, \\ &\qquad \forall k,h \in \{1,2,\cdots,M\}; \end{aligned} \quad (37)$$

$$\begin{aligned} &\mathbb{P}_m^{(n)}(j|i) \\ &= \frac{\sum_{t=0}^{T_c-2}\mathbf{1}\left(m_t \in \mathcal{T}_m(i), m_{t+1} \in \mathcal{T}_m(j)\right)}{\sum_{t=0}^{T_c-1}\mathbf{1}\left(m_t \in \mathcal{T}_m(i)\right)}, \\ &\qquad \forall i,j \in \{1,2,\cdots,N\}; \end{aligned} \quad (38)$$

$$\begin{aligned} &\text{and } \mathbb{P}_{gm}^{(n)}\left((h,j)|(k,i)\right) \\ &= \frac{\sum_{t=0}^{T_c-2}\mathbf{1}_{gm}}{\sum_{t=0}^{T_c-1}\mathbf{1}\left(\|\mathbf{g}_t\|^2 \in \mathcal{T}_g(k), m_t \in \mathcal{T}_m(i)\right)}, \\ &\forall k,h \in \{1,2,\cdots,M\}, \forall i,j \in \{1,2,\cdots,N\}, \end{aligned} \quad (39)$$

where in (39) we have defined

$$\begin{aligned} \mathbf{1}_{gm} \triangleq \mathbf{1}\Big(&\|\mathbf{g}_t\|^2 \in \mathcal{T}_g(k), m_t \in \mathcal{T}_m(i), \\ &\|\mathbf{g}_{t+1}\|^2 \in \mathcal{T}_g(h), m_{t+1} \in \mathcal{T}_m(j)\Big). \end{aligned}$$

c) $\mathbb{P}_g \leftarrow \mathbb{P}_g + \mathbb{P}_g^{(n)}$, $\mathbb{P}_m \leftarrow \mathbb{P}_m + \mathbb{P}_m^{(n)}$, $\mathbb{P}_{gm} \leftarrow \mathbb{P}_{gm} + \mathbb{P}_{gm}^{(n)}$.

d) $n \leftarrow n+1$.

3: Repeat Step 2 until n equals N_i, which is the pre-defined number of iterations.

4: $\mathbb{P}_g \leftarrow \mathbb{P}_g/N_i$, $\mathbb{P}_m \leftarrow \mathbb{P}_m/N_i$, and $\mathbb{P}_{gm} \leftarrow \mathbb{P}_{gm}/N_i$.

References

[1] L. C. Godara, "Application of antenna arrays to mobile communications, part II: beam-forming and direction-of-arrival considerations," *Proc. IEEE*, vol. 85, pp. 1195–1245, Aug. 1997.

[2] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 1341–1365, Oct. 2008.

[3] D. Love, R. Heath, Jr, and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2735–2747, Oct. 2003.

[4] K. Huang, V. Lau, and D. Kim, "Event-driven optimal feedback control for multiantenna beamforming," *IEEE Trans. Signal Process.*, vol. 58, pp. 3298–3312, Jun. 2010.

[5] W. Santipach and M. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," in *Proc. IEEE Mil. Comm. Conf.*, 2003.

[6] J. Roh and B. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: a VQ-based approach," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1101–1112, Mar. 2006.

[7] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.

[8] K. Huang, R. Heath, Jr, and J. Andrews, "SDMA with a sum feedback rate constraint," *IEEE Trans. Signal Process.*, vol. 55, pp. 3879–3891, Jul. 2007.

[9] C. Swannack and G. Uysal-Biyikoglu, "MIMO broadcast scheduling with quantized channel state information," in *Proc., IEEE Int. Symp. Inf. Theory*, 2006.

[10] S. Sanayei and A. Nosratinia, "Exploiting multiuser diversity with only 1-bit feedback," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2005.

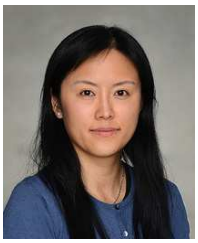
[11] M. Medard, I. Abou-Faycal, and U. Madhow, "Adaptive coding with pilot signals," in *Proc. 38th Annual Allerton Conf. on Communication, Control and Computing.*, Allerton, IL, Oct. 2000.

[12] M. Dong, L. Tong, and B. Sadler, "Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels," *IEEE Trans. Signal Processing*, vol. 52, pp. 1403–1418, May 2004.

[13] T. Kim, D. Love, and B. Clerckx, "MIMO systems with limited

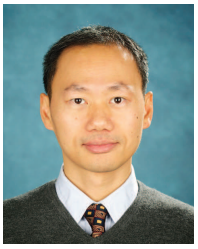
rate differential feedback in slowly varying channels," *IEEE Trans. Commun.*, vol. 59, pp. 1175–1189, Apr. 2011.
[14] J. Choi, B. Clerckx, N. Lee, and G. Kim, "A new design of polar-cap differential codebook for temporally/spatially correlated MISO channels," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 703–711, Feb. 2012.
[15] P, Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels: a survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, pp. 57–80, Sep. 2008.
[16] C. Tan and N. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun.*, vol. 48, pp. 2032–2040, Dec. 2000.
[17] Q. Zhang and S. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, pp. 1688–1692, Nov. 1999.
[18] B. Banister and J. Zeidler, "A simple gradient sign algorithm for transmit antenna weight adaptation with feedback," *IEEE Trans. Signal Processing*, vol. 51, pp. 1156–1171, May 2003.
[19] K. Huang, R. W. Heath, Jr., and J. Andrews, "Limited feedback beamforming over temporally-correlated channels," *IEEE Trans. Signal Process.*, vol. 57, pp. 1959–1975, May 2009.
[20] C. Simon and G. Leus, "Feedback quantization for linear precoded spatial multiplexing," *EURASIP J Adv Signal Process*, pp. 1–13, 2008.
[21] D. Sacristán-Murga and A. Pascual-Iserte, "Differential feedback of MIMO channel gram matrices based on geodesic curves," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3714–3727, Dec. 2010.
[22] J. Yang and D. Williams, "Transmission subspace tracking for MIMO systems with low-rate feedback," *IEEE Trans. Commun.*, vol. 55, pp. 1629–1639, Aug. 2007.
[23] C. Yeung and D. Love, "Optimization and tradeoff analysis of two-way limited feedback beamforming systems," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 2570–2579, May 2009.
[24] D. Love, "Duplex distortion models for limited feedback MIMO communication," *IEEE Trans. Signal Process.*, vol. 54, pp. 766–774, Feb. 2006.
[25] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2005.
[26] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. A Bradford Book, 1998.
[27] A. Barto, S. Bradtke, and S. Singh, "Learning to act using real-time dynamic programming," *Artif. Intell.*, vol. 72, pp. 81–138, 1995.
[28] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: a survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.
[29] J. Shao, *Mathematical Statistics*. Springer, 2003.
[30] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Springer, 1992.
[31] K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2562–2579, Oct. 2003.
[32] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.
[33] A. Barto and S. Singh, "On the computational economics of reinforcement learning," in *Proc. of the 1990 Connectionist Models Summer School*, 1990.
[34] 3GPP TS 36.300, V.11.9.0 Release 11, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Overall description," 2014.

Sun Sun (S'11) received the B.S. degree in Electrical Engineering and Automation from Tongji University, Shanghai, China, in 2005. From 2006 to 2008, she was a software engineer in the Department of GSM Base Transceiver Station of Huawei Technologies Co. Ltd.. She received the M.Sc. degree in Electrical and Computer Engineering from University of Alberta, Edmonton, Canada, in 2011. Now, she is pursuing her Ph.D. degree in the Department of Electrical and Computer Engineering of University of Toronto, Toronto, Canada. She is interested in the areas of statistical learning, stochastic optimization, distributed control, and network resource management.



Min Dong (S'00-M'05-SM'09) received the B.Eng. degree from Tsinghua University, Beijing, China, in 1998, and the Ph.D. degree in electrical and computer engineering with minor in applied mathematics from Cornell University, Ithaca, NY, in 2004. From 2004 to 2008, she was with Corporate Research and Development, Qualcomm Inc., San Diego, CA. In 2008, she joined the Department of Electrical, Computer and Software Engineering at University of Ontario Institute of Technology, Ontario, Canada, where she is currently an Associate Professor. She also holds a status-only Associate Professor appointment with the Department of Electrical and Computer Engineering, University of Toronto since 2009. Her research interests are in the areas of statistical signal processing for communication networks, cooperative communications and networking techniques, and stochastic network optimization in dynamic networks and systems.

Dr. Dong received the Early Researcher Award from Ontario Ministry of Research and Innovation in 2012, the Best Paper Award at IEEE ICCC in 2012, and the 2004 IEEE Signal Processing Society Best Paper Award. She was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS during 2009-2013, and currently serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She has been an elected member of IEEE Signal Processing Society Signal Processing for Communications and Networking (SP-COM) Technical Committee since 2013.



Ben Liang (S'94-M'01-SM'06) received honors-simultaneous B.Sc. (valedictorian) and M.Sc. degrees in Electrical Engineering from Polytechnic University in Brooklyn, New York, in 1997 and the Ph.D. degree in Electrical Engineering with Computer Science minor from Cornell University in Ithaca, New York, in 2001. In the 2001 - 2002 academic year, he was a visiting lecturer and post-doctoral research associate at Cornell University. He joined the Department of Electrical and Computer Engineering at the University of Toronto in 2002, where he is now a Professor. His current research interests are in mobile communications and networked systems. He has served as an editor for the IEEE Transactions on Wireless Communications and an associate editor for the Wiley Security and Communication Networks journal, in addition to regularly serving on the organizational or technical committee of a number of conferences. He is a senior member of IEEE and a member of ACM and Tau Beta Pi.