Q-Learning Algorithms for Constrained Markov Decision Processes with Randomized Monotone Policies: Application to MIMO Transmission Control

Dejan V. Djonin, Vikram Krishnamurthy, *Fellow, IEEE*

***Abstract*— This paper presents novel Q-learning based stochastic control algorithms for rate and power control in V-BLAST transmission systems. The algorithms exploit the supermodularity and monotonic structure results derived in the companion paper. Rate and power control problem is posed as a stochastic optimization problem with the goal of minimizing the average transmission power under the constraint on the average delay that can be interpreted as the Quality of Service (QoS) requirement of a given application. Standard Q-learning algorithm is modified to handle the constraints so that it can adaptively learn structured optimal policy for unknown channel/traffic statistics. We discuss the convergence of the proposed algorithms and explore their properties in simulations. To address the issue of unknown transmission costs in an unknown time-varying environment, we propose the variant of Q-learning algorithm in which power costs are estimated in on-line fashion, and we show that this algorithm converges to the optimal solution as long as the power cost estimates are asymptotically unbiased.**

***Index Terms*— Q learning, Supermodularity, Monotone Policies, Randomized Policies, Constrained Markov Decision Process, Transmission Scheduling, V-BLAST, Delay Constraints, Reinforcement Learning**

I. Introduction

This paper addresses the problem of structured learning of rate and power control policy for transmission over wireless Multiple Input Multiple Output (MIMO) channel and under the constraint on transmission latency. Several structural results on the optimal costs and policies have been derived in the companion paper [1]. It has been shown in [1] that the optimal rate allocation action is monotonic increasing in the buffer occupancy and that control policy optimization can be divided into two separate problems of low-layer bit-loading and high-layer total rate allocation. In this paper, we exploit these structural results to derive computationally efficient stochastic control algorithms.

Summary of the Contributions: The most important contributions of this paper are:

- Application of online policy learning algorithms for the computation of the optimal rate scheduling algorithms for delay-constrained V-BLAST transmission in imperfectly known channel and traffic environments with simulated costs.
- Utilizing structural results on the optimal rate scheduling policy with the goal of improving the convergence rate of online learning algorithms.
- Analytical formulation and numerical examination of three novel algorithms designed to incorporate submodular linear constraint in the standard Q-learning algorithm to improve its convergence.

The main ideas of this paper are novel from a communications perspective as well as learning based perspective.

Communications Perspective: The problem addressed in this paper is a cross-layer optimization problem as we jointly consider the statistics of the traffic arriving into the adaptive V-BLAST transmitter and the transmitter adaptation based on the channel statistics. The optimization goal is to reduce the total transmission power over all transmitter antennas, while maintaining the constraint on transmission delay satisfied. Due to the consideration of the transmit buffer, this problem is inherently a dynamic stochastic optimization problem and can be stated as a Constrained Markov Decision Process (CMDP).

The problem of transmission control optimization for single-channel systems, with the objective of average power or bit-error rate minimization under the latency constraints, has been previously addressed in [2], [3]. Multichannel rate adaptation for (Orthogonal Frequency Division Multiplexing) OFDM systems with delay constraint has been addressed in [4]. However, none of these results discuss the problem of transmission control adaptation when channel and/or traffic statistics is unknown. Here, we address this problem and present several algorithms for adaptive transmission control. These adaptive learning algorithms are based on the ideas of stochastic approximation and reinforcement learning [5]. To the best knowledge of the authors, the structured submodular Q-learning algorithm proposed in this paper is also novel from the control-theoretic viewpoint.

From a signal transmission perspective, the importance of the addressed problem is threefold: (1) We address the adaptive policy learning, as the wireless channels and traffic statistics are usually not a priori known. Thereby the static optimization is not suitable to address the transmission control problem, (2) We address the MIMO channels as they provide higher capacities than single channel systems [6], [7]. The MIMO channel capacity can be further increased by employing per-antenna power and rate allocation at the transmitter (e.g. see power adaptation for Bell-Labs layered space-time (V-BLAST) addressed in [8], [9]), (3) We incorporate the consideration of real-time traffic in order to reduce the transmission latency and satisfy different user QoS conditions.

Learning-based Perspective: The problem of policy learning for the analyzed V-BLAST power and rate adaptation can be addressed either using discrete stochastic approximation algorithms by learning the control policy directly (see [10] and references within) or by using continuous stochastic approximation algorithms such as Q-learning [5], actor-critic

The research on this paper has been supported by the NSERC PostDoctoral Fellowship Award and the NSERC strategic grant. The authors are with the Department of Electrical and Computer Engineering, 2356 Main Mall, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada, Phone: +1 604 827 3268, +1 604 822 2653, Fax: +1 604 822 5949, E-mails: ddjonin@ece.ubc.ca; vikramk@ece.ubc.ca.

methods [11] and policy space methods [12]. In this paper we have decided to pursue the continuous stochastic approximation approach as the discrete version would involve the search for the optimal policy within a very large set of optimal policies. Traditionally, both discrete stochastic approximation and Q-learning were used for unconstrained Markov Decision Process (MDP) and calculation of pure optimal policies. Due to equivalence in costs of CMDP and a Lagrangian MDP formulation of a CMDP [13], Q-learning can be applied to compute the optimal pure policy for a fixed Lagrangian multiplier and active constraint. The optimal randomized policy of a CMDP can be computed as a mixed policy of two optimal pure policies for two different Lagrangian multipliers. We also present an iterative algorithm to find these Lagrangian multipliers and compute optimal randomized policies.

Each of the pure policies that constitute the optimal randomized policy possesses a known structure, that is, rate allocation actions are monotonically increasing with the buffer state. This implies that Q-factors possess a submodularity property that can be stated as a linear constraint on Q-factors and easily utilized in Q-learning algorithm[1]. The rationale is that by imposing submodularity structure on Q-factors, Q-learning more quickly searches through the policies and avoids considering non-structured policies that are known to be non-optimal. It has been shown in [1] that reduction in policy space, achieved by considering only structured policies, can be several orders of magnitude. Further, unlike actor-critic methods [11], Q-learning algorithm has well-explored convergence properties that can be shown to carry over to the structured version of Q-learning.

In practice, costs of such CMDP can be estimated online during the learning phase and sampled costs can be used to update the Q-factors. Q-learning converges to the optimal solution with probability one as long as the cost estimates are asymptotically unbiased. We discuss how to perform power cost estimation in case that powers are adapted at a faster rate than the transmission rates. This approach has an added advantage that transmission adaptation actions (that have to be negotiated between the transmitter and the receiver) can be performed less frequently than the power control actions. Furthermore, rate control actions can be based on a more coarse quantization of the channel state than the power control actions. This results in a more efficient Q-learning algorithm.

Paper outline: The outline of the paper is as follows. We formulate the V-BLAST power and rate control problem using stochastic control framework and Constrained Markov Decision Processes in Section II. Respective costs and transition probabilities are identified for such a problem in Section III. Section IV presents a summary of structural properties of optimal policies. In Section IV we utilize this structure of the optimal policy and propose several methods to improve the convergence rate of the Q learning algorithm. This approach results in novel structured Q-learning algorithms for CMDPs that are posed as stochastic constrained optimization problem with linear constraints. We propose three algorithms to solve that constrained optimization problem. Namely, we address the primal-dual algorithm, primal projection algorithm and the submodular parameterization algorithm. In Section VI, we numerically explore performances of the proposed structured Q-learning algorithms for delay-constrained V-BLAST rate and power control. The simulations show that primal projection method best utilizes a priori known structure of the optimal policy for both stringent and relaxed delay constraints.

[1]As opposed to the presented structured Q-learning algorithm, it is difficult to incorporate submodular constraints and ensure convergence to the optimal discrete policy using policy space search methods discussed in [12].

II. V-BLAST Transmission Model

Notation: A discrete-time slotted model is used throughout the paper. A time slot n is defined as the time interval $[nT, (n+1)T)$ and controller decision in this time slot is made at the beginning of that interval at time nT. Let $x^{(n)}$ denote the discrete-time (in general random) variable x at time slot n. To avoid cumbersome notation, we will drop the time-slot superscript designation whenever that does not cause confusion. Let $|\mathcal{C}|$ denote the cardinality of a certain finite set $\mathcal{C}$, and $\mathbb{P}[\cdot]$ denote the probability measure. Let $\mathcal{N}_0$ be the set of integers including 0.

Fig. 1 shows a schematic representation of the V-BLAST transmitter and receiver model used in this paper. The transmitter is equipped with the transmission buffer of length L. The task of the controller is to choose rates and powers for each of the t transmission antennas. Let us denote the buffer occupancy in bits at the beginning of the n-th time slot with $b^{(n)}$, where $b^{(n)} \in \mathcal{B}$ and the buffer state space is $\mathcal{B} = \{0, 1, 2, \ldots, L\}$.

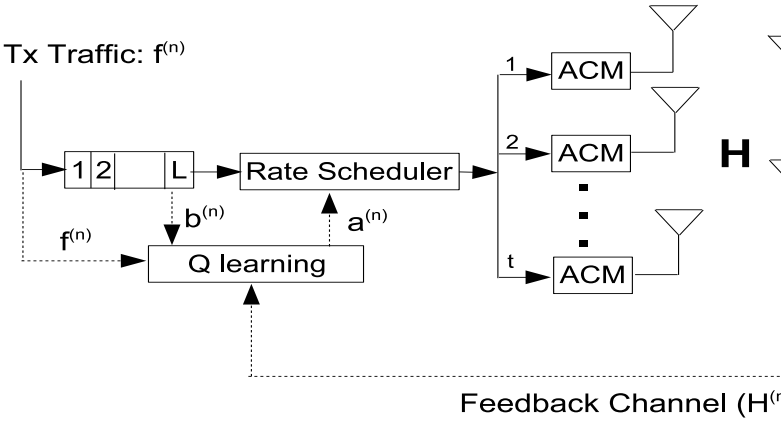

Fig. 1. MIMO Transmission Model with Q-learning algorithm. ACM stands for adaptive coding and modulation.

The transmission buffer is continually supplied with the incoming traffic from a higher layer application. Let $f^{(n)}$ be the number of packets stored into the transmission buffer during the n-th time slot. It is assumed that for all n, $f^{(n)}$ is the element of a finite state space $\mathcal{F} = \{0, 1, \ldots, F\}$ and the packet length is G bits. Furthermore, we assume the following:

In our results to follow we will also use the following alternative assumption that completely omits the use of flow controller will be used.

A 1: The Number of packets stored into the transmission buffer is an ergodic Markov Chain with transition probabilities $p^f(f^{(n+1)}|f^{(n)})$ that are independent of the chosen action, buffer occupancy and channel state.

Consider that traffic $e^{(n)}$ arriving onto the transmission buffer is Markovian and independent of the buffer occupancy and actions taken. Out of these $e^{(n)}$ packets, only $f^{(n)}$ packets are stored into the transmission buffer. Therefore, A1 can not be satisfied if number of packets $f^{(n)}$ that is stored into the transmission buffer is dependent on the buffer occupancy or actions, i.e. A1 can not be satisfied if there are buffer overflows in the finite transmission buffer.

Markov model for the incoming traffic is sufficient since the incoming traffic state space $\mathcal{F}$ is finite and control decisions are made periodically at the end of each time slot. Therefore, it is not necessary to consider the semi-Markov process for the incoming traffic.

MIMO channel considered in this paper is a point-to-point wireless channel with t transmit and r receive antennas, that satisfy $r \geq t$. The channel is considered to be block fading and constant during a time slot of length T. Furthermore, at time slot n, the MIMO channel is completely described with the complex $r \times t$ dimensional channel matrix $\mathbf{H}^{(n)}$ containing the elements $h_{i,j}$.

Let $\mathbf{d}^{(n)} = [d_1, \ldots, d_t]^T$ be the vector of transmitted symbols employing certain modulation format, from all of the t transmit antennas. Each of the transmitted data streams over t antennas can contain independent information. Then, received signal vector $\mathbf{y}^{(n)} = [y_1, \ldots, y_r]^T$ can be presented in the following complex baseband vector form

$$\mathbf{y}^{(n)} = \mathbf{H}^{(n)}\mathbf{d}^{(n)} + \mathbf{w}^{(n)} \tag{1}$$

while $\mathbf{w}^{(n)} = [w_1, \ldots, w_r]^T$ is the noise vector. The elements of $\mathbf{w}$ are assumed to be independent and identical distributed (i.i.d) Gaussian random variables with zero mean and variance σ^2. Channel matrix $\mathbf{H}^{(n)}$ is assumed to be dependent only on the previous time slot i.e. $\mathbb{P}[\mathbf{H}^{(n)}|\mathbf{H}^{(n-1)}, \mathbf{H}^{(n-2)}, \ldots] = \mathbb{P}[\mathbf{H}^{(n)}|\mathbf{H}^{(n-1)}]$ and the sequence of channel matrices $\mathbf{H}^{(n)}, n = 1, 2, \ldots$ constitutes a continuous value Markov process.

A. Receiver Structure

In the above MIMO channel, signals from all of the t antennas are received on all of the r receiver antennas. To recover and estimate the transmitted signals, several receiver structures have been devised. These include the linear receivers such as zero-forcing and Minimum Mean Square Estimation (MMSE) receivers, and non-linear successive interference cancelation receivers. Each of these receivers compute estimates for all of the t independent data streams.

Next we consider the zero-forcing (ZF) linear detector and show that by employing this detector, the MIMO channel is decoupled into t parallel independent channels[2]. The zero forcing detector assumes that knowledge of channel gains $h_{i,j}$ is known at the receiver. In this receiver, the received signal $\mathbf{y}$ at time slot n is multiplied by the pseudoinverse $\mathbf{H}^{\dagger}$. The post-detection SNR, normalized by the nominal transmission power of $P_o = 1mW$, and associated with k-th transmission antenna when linear ZF equalizer is used can be expressed as

$$\gamma_k = \frac{\gamma_0}{[\mathbf{H}^{\dagger}\mathbf{H}]^{-1}_{kk}}, \qquad k = 1, \ldots, t \tag{2}$$

where γ_0 is the normalized received SNR at each receive antenna and is defined as $\gamma_0 = 1/\sigma^2$.

We will assume that quantized information on the post-detection SNR is provided to the transmitter, and that this information is utilized in the controller to choose the current rates and power levels in all of the t transmit antennas. For the k-th transmit antenna ($k = 1, 2, \ldots, t$), we assume that post-detection SNR γ_k is quantized using thresholds at $\{\Gamma_{k0}, \Gamma_{k1}, \ldots, \Gamma_{kK}\}$, where $\Gamma_{k0} = 0$ and $\Gamma_{kK} = \infty$. Denote with $\mathcal{H}^s$ the set of all quantized channel states corresponding to a certain transmit antenna. Therefore, there will be K channel states in $\mathcal{H}^s$ for each of the single antenna post-detection SNR-s. Let the current channel state associated with the k-th transmitter ($k = 0, 1, \ldots, t$) be denoted with h_k and $h_k \in \mathcal{H}^s$. Let us denote with $\mathcal{H} = \mathcal{H}^s \times \mathcal{H}^s \times \cdots \times \mathcal{H}^s$ the composite MIMO channel state defined as the Cartesian product of state spaces $\mathcal{H}^s$ of quantized post-detection SNR-s for each of t transmitter data streams. The composite channel state of all transmit channels is denoted with $h = \{h_1, h_2, \ldots, h_t\} \in \mathcal{H}$ and we will adopt the following assumption regarding its statistical evolution:

A 2: The sequence of channel states, forms an ergodic first order Markov chain with transition probabilities $p^h(h^{(n+1)}|h^{(n)})$ and is independent of the action, buffer state and incoming traffic state.

III. V-BLAST Power and Rate Control Problem as CMDP

This section provides a detailed formulation of the V-BLAST Power and Rate Control Problem (V-BLAST-PRCP) formulated as CMDP. A CMDP is completely described through its state space, action space, transition probabilities and cost criteria. Let $\mathcal{S}$ denote an arbitrary finite set called the *state space* while $\mathcal{A}$ denote the finite set called the *action set* of a CMDP. The proceeding definitions are foundation building blocks of the Q-learning algorithm to be discussed in Section V.

State Space: Utilizing definitions of Section II the state space $\mathcal{S}$ of V-BLAST-PRCP is the composite space comprising of buffer space $\mathcal{B}$, incoming traffic space $\mathcal{F}$ and the channel state space $\mathcal{H}$, i.e. $\mathcal{S} = \mathcal{H} \times \mathcal{B} \times \mathcal{F}$ where $\times$ denotes the Cartesian product.

Action Space: The action in the V-BLAST-PRCP is interpreted as the composite rate allocation of the individual transmitter antennas. Let $a_l \in \mathcal{A}^s$ denote the number of bits that are allocated to antenna l where $\mathcal{A}^s$ is the set of all possible single-antenna bit allocations. As shown in Fig. 1 these a_l bits are processed by the adaptive coding and modulation (ACM) block to produce N consecutive symbols

[2]ZF detector has been used in the simulations of Section VI. However, the CMDP model, proposed adaptive algorithm and utilized structural results are also valid for MMSE detectors with appropriate change in the power costs.

transmitted from antenna $l = 1, 2, \ldots, t$. Let the composite action $a = \{a_1, a_2, \ldots, a_t\}$ denote the bit allocations across all t transmit antennas. The set of composite actions is equal to $\mathcal{A} = \mathcal{A}^s \times \mathcal{A}^s \times \cdots \times \mathcal{A}^s$. Let us also define function $\Psi(a), a \in \mathcal{A}$ which returns the number of bits retrieved from the buffer if action a is applied, i.e.

$$\Psi(a) = \sum_{i=1}^{t} a_i. \tag{3}$$

Define the set of Markovian *admissible* policies $\Phi = \{a = \{a^{(n)}\} | a^{(n)}$ is measurable w.r.t. $\Omega^{(n)}, \forall n = 0, 1, \ldots\}$. Let $\Omega^{(n)}, n \geq 0$ denote the σ-algebra generated by the observed system state $s^{(0)}, \ldots, s^{(n)}$ at time n. This means that $a^{(n)}$ is a (potentially) random function of current state $s^{(n)}$. Let Φ_D denote the set of all pure policies where $a^{(n)}$ is a deterministic function of current state $s^{(n)}$.

We now introduce the following unichain assumption on the set of optimal policies Φ:

A 3: The set of admissible policies Φ for the RFCP CMDP comprises of unichain policies.

This assumption establishes regularity conditions of the CMDP that ensures the existence of the optimal policy for the average cost problems (for more details see [14]). A CMDP is *unichain* [14] if every policy where $a^{(n)}$ is a deterministic function of $s^{(0)}, \ldots, s^{(n)}$ induces a single recurrent class plus possibly an empty set of transient states.

Transition Probabilities: When the system is in state $s \in \mathcal{S}$, a finite number of possible actions which are elements of the set $\mathcal{A}$ can be taken. Let $a^{(n)}$ denote the action taken by the decision maker at the time n. For a given policy, the evolution of a MDP is Markovian with transition probabilities

$$p(s_l | s_j, a) = \mathbb{P}[s^{(n+1)} = s_l | s^{(n)} = s_j, a^{(n)} = a] \tag{4}$$

for some $s_l, s_j \in \mathcal{S}$, $a \in \mathcal{A}_{s_j}$ and $n = 0, 1, \ldots$.

Based on Assumptions 1 and 2, the transition probability of V-BLAST-PRCP between the composite state $s = \{b, h, f\}$ and $s' = \{b', h', f'\}, s, s' \in \mathcal{S}$ when action a is taken is given with

$$p(s' | s, a) = p^h(h' | h) p^f(f' | f) I_{\{b' = \min(b - \Psi(a) + Gf', L)\}}. \tag{5}$$

where $I_{\{l\}}$ is the indicator function that returns 1 if l is true and 0 otherwise.

Cost Criteria: We will adopt the average expected cost as the optimization criteria in V-BLAST-PRCP. For any admissible policy $\pi \in \Phi$, let the infinite horizon cost conditioned on initial state $s^{(0)}$ be defined as

$$C_{s^{(0)}}(\pi) = \mathbb{E}_\pi \left[\lim_{N \to \infty} \sup \frac{1}{N} \sum_{i=1}^{N} c(s^{(n)}, a^{(n)}) | s^{(0)} \right] \tag{6}$$

where the expectation is over randomized actions $a^{(n)}$ and system state $s^{(n)}$ evolution for $n = 1, 2, \ldots$. The goal is to compute the optimal policy π^* that minimizes the cost (6)

$$C^*_{s^{(0)}} = \inf_{\pi \in \Phi} C_{s^{(0)}}(\pi), \tag{7}$$

subject to the global constraint

$$D_{s^{(0)}}(\pi) = \mathbb{E}_\pi \left[\lim_{N \to \infty} \sup \frac{1}{N} \sum_{i=1}^{N} d(s^{(n)}, a^{(n)}) | s^{(0)} \right] \leq \tilde{D}. \tag{8}$$

Let finite cost $c(s^{(n)}, a^{(n)}) \geq 0$ be the instantaneous cost of taking action $a^{(n)}$ in the state $s^{(n)}$. For any linear V-BLAST receiver the power cost for the composite channel state $h = \{h_1, h_2, \ldots, h_t\}$ and composite rate action $a = \{a_1, a_2, \ldots, a_t\}$ can be expressed as the total power necessary for transmission with a given average bit-error rate, i.e.

$$c([h, b, f], a) = \sum_{i=1}^{t} P^i(h_i, a_i). \tag{9}$$

where $P^i(h_i, a_i)$ is a single-channel power needed to transmit with rate action a_i over a channel state h_i. Let the instantaneous delay cost be defined as

$$d([h, b, f], a) = \frac{b}{G\bar{F}} \tag{10}$$

where $\bar{F}$ is the average number of incoming packets in a time slot and G is the length of each packet in bits. For $d(s, a)$ given above and according to the Little's formula, (8) describes the constrained average delay incurred in the buffer. Constraint cost $\tilde{D} \geq 0$ is a user specified parameter. Any policy π^* that minimizes $C_{s^{(0)}}(\pi)$ will be called the optimal policy. The cost of the policy π^* that is optimal subject to constraint (8) will be denoted by $C^*(\tilde{D})$.

Single-channel power cost $P^i(h_i, a_i)$ when action a_i is applied, is a random variable dependent on the random post-detection SNR $\gamma \in \Gamma_{h_i}$ conditioned on channel state h_i. However, as is known from [15], the equivalent immediate costs in the case of random immediate costs can be calculated as the average cost for a given state-action pair. Therefore

$$P^i(h_i, a_i) = \int_{\Gamma_{i(h_i - 1)}}^{\Gamma_{ih_i}} P(\gamma, a_i) p^{h_i}(\gamma) d\gamma \tag{11}$$

where $P(\gamma, a)$ is the power needed to transmit with rate a over a channel with SNR γ with a given bit-error rate of BER_t. The expectation is over signal to noise ratio (SNR) γ conditioned on the channel state being in state h_i i.e. $\gamma \in [\Gamma_{i(h_i - 1)}, \Gamma_{ih_i})$. [3]

Furthermore, $P(\gamma, a)$ can be calculated from

$$BER_t = BER(\gamma P(\gamma, a), a) \tag{12}$$

for given BER_t. The expression for the bit-error rate $BER(g, a)$ is a function of the instantaneous signal to noise ratio g and the rate action a and depends on the utilized modulation format. In the numerical results Section VI we use the uncoded M-ary quadrature modulation (QAM) in each of the transmission antennas and its bit-error rate expression will be approximated with (see e.g. [16])

$$BER(g, a) = 0.2 \times \exp\left[\frac{-1.6g}{(2^v - 1)} \right]. \tag{13}$$

[3] As an alternative to the above calculation of power costs, in Section V-C, we will discuss online estimation of power costs that can be used in conjunction with the online Q-learning algorithm.

Therefore, for uncoded MQAM the immediate single-channel power cost for channel SNR γ can be expressed as

$$P(\gamma, a) = \frac{-0.625 \log(5BER_t)}{\gamma}(2^a - 1). \quad (14)$$

Let $c(s, a; \lambda)$ be the Lagrangian cost given with

$$c(s, a; \lambda) = c(s, a) + \lambda d(s, a) \quad (15)$$

for a certain Lagrangian multiplier $\lambda > 0$. As discussed in [13] and [1] the optimal policy of the CMDP with one constraint is a mixture of two pure policies that are optimal for unconstrained MDP with costs given as in (15) and two different Lagrangian multipliers.

IV. Summary of Structural Results on Optimal Policies

In this section we review two theorems, whose proofs are given in the companion paper [1]. These results will allow us to simplify the computational complexity and exploit the structural results of the optimal policy in Q-learning algorithms that are to be discussed in Section V.

Action Reduction: We first note that the action space of the V-BLAST-PRCP is of dimensions $|\mathcal{A}^s|^t$ that can be very large for large transmit antenna arrays. The following theorem demonstrates that the action space can be reduced to the set with on the order of $t|\mathcal{A}^s|$ states and that the optimal policy of the V-BLAST-PRCP will only utilize actions from this reduced action set.

Theorem 1: For the V-BLAST-PRCP, if the transmission cost $c([h, b, f], a)$ have the form (9), then the composite action set $\mathcal{A}$ containing $(\alpha + 1)^t$ actions can be exponentially decreased in cardinality to the reduced action set $\tilde{\mathcal{A}}$ with $t\alpha+1$ actions. ∘

To compute the optimal policy of V-BLAST-PRCP, for a certain channel state h and traffic state f, only the action $a(u)$ that has the minimal transmission cost among the actions that retrieve a fixed amount of data $\Psi(a) = u$ from the transmission buffer should be considered i.e.

$$a(u) = \arg \min_{\{a|\Psi(a)=u\}} c([h, b, f], a). \quad (16)$$

All other actions can be dropped from the model. To utilize the results of the previous theorem, we can define the reduced action set as $\tilde{A} = \{a(r)|r = 0, \ldots, t \max(A^s)\}$. Using the reduced actions sets, in the proceeding sections we assume for simplicity that buffer retrieval function $\Psi(a) = a$ is equal to the ordinal number of action a from the reduced action set $\tilde{A}$. Let $A = |\tilde{A}|$ be the number of actions in the reduced action set.

Monotonic Policies: The Q-learning algorithm is based on the adaptive iterative learning of Q factors of a Markov Decision Process. Q factors are defined as

$$Q(s, a; \lambda) = \left[\sum_{s'} p(s'|s, a) \left(c(s, a; \lambda) + V(s'; \lambda)\right)\right] \quad (17)$$

where $V(s; \lambda)$ denotes the value function of a certain state $s \in \mathcal{S}$ that is the solution of the Bellman's equation

$$V(s; \lambda) = \min_a \left[\sum_{s'} p(s'|s, a) \left(c(s, a; \lambda) + V(s'; \lambda)\right)\right] \quad (18)$$

for a fixed Lagrange multiplier λ. Function $Q : A \times \mathcal{B} \times \mathcal{H} \times \mathcal{F} \rightarrow \Re$ is called submodular (has decreasing differences) in (a, b) for a fixed parameters $h \in \mathcal{H}$ and $f \in \mathcal{F}$, if for all $a' \geq a$ and $b' \geq b$,

$$Q(a', b'; h, f) - Q(a, b'; h, f) \leq Q(a', b; h, f) - Q(a, b; h, f). \quad (19)$$

It has been shown in [17] that if $Q(s, a)$ is submodular in (s, a) then the optimal action $\max_s Q(s, a)$ of the MDP for certain state s is monotonically increasing in the state s.

The following assumption and definitions for the stated transmission control problem will be used to establish the below result on monotonic policies.

A 4: Set of feasible actions $\mathcal{A}_s$ in state $s = [h, b, f] \in \mathcal{S}$ is a non-empty set of actions $a \in \mathcal{A}$ for which $b + Gf' - \Psi(a) \leq L$ and $b - \Psi(a) \geq 0$ and any $f' \in \mathcal{F}$.

This assumption states that there exist such a feasible policy that will not lead to transmit buffer overflows.

Definition 1: For any $0 \leq q \leq 1$, *mixed policy* π is a randomized policy formed of two pure policies π_1 and π_2 such that policy π_1 is applied with probability q and policy π_2 is applied with probability $1 - q$.

The next concept we will use is multimodularity. Multimodularity extends the convexity property of continuous functions defined on Euclidean space to real-valued functions defined on a discrete set. We will call $\mathcal{M} = \{[-1, 0], [1, -1], [0, 1]\}$ a 2-dimensional multimodular base and let $\mathcal{X}$ be a convex subset of the set of ordered pairs of integers $\mathcal{Z}^2$ [18].

Definition 2: (Multimodularity) A real-valued function $f : \mathcal{X} \rightarrow \Re$ is multimodular with respect to base $\mathcal{M}$ if for all $x \in \mathcal{X}$, and $v, w \in \mathcal{M}$, $v \neq w$ the following holds:

$$f(x + v) + f(x + w) \geq f(x) + f(x + v + w). \quad (20)$$

We will use the multimodularity property of discrete functions due to its property that it remains preserved after minimization over any subset of parameters of a multimodular function (see key Lemma 61 in [18]).

According to the previous definition, mixed policy π is a randomized policy that is convex combination of pure policies π_1 and π_2.

Definition 3: Pure policy π is *non-decreasing in the buffer state* b if the ordinal number (index) of the action $a = \pi([h, b, f])$ taken in state $[h, b, f]$ is non-decreasing in buffer state b for each channel state h and traffic state f.

Theorem 2: Consider V-BLAST-PRCP defined in Section III. Let assumptions A2, A1, A3 hold. Furthermore, assume that the following assumptions holds

A 5: Lagrangian cost $c(s, a; \lambda) = c([h, b, f], a; \lambda)$ defined in (15) be multimodular function of $b, -a$, submodular function of b, a for any $f \in \mathcal{F}$, $\lambda \in \Re^+$, $h \in \mathcal{H}$.

Then for cost constraint $\tilde{D} > 0$, the optimal randomized policy $\pi^*([h, b, f])$ is a mixed policy of two pure policies $\pi^1([h, b, f])$ and $\pi^2([h, b, f])$. Both $\pi^1([h, b, f])$ and