Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Structural Results on the Optimal Transmission Scheduling Policies and Costs for Correlated Sources and Channels

Dejan V. Djonin and Vikram Krishnamurthy

Abstract— **The problem of transmission adaptation over a correlated time-varying wireless channel is formulated as a Constrained Markov Decision Process. The model includes a transmission buffer and finite state Markov model for time-varying radio channel and incoming traffic. This cross-layer optimization problem is formulated as to minimize the transmission cost (e.g. power or bit-error-rate) under the constraint on a buffer cost such as the transmission delay. Under the assumptions on submodularity and convexity of the cost function it is shown that the optimal randomized policy is monotonically increasing with the increase of the buffer state.**

Index Terms— **Value function, scheduling, optimal policy, Markov Decision process, correlated sources, correlated channels, transmission adaptation, supermodularity, stochastic dominance**

I. Introduction

The use of adaptation of transmission parameters over time-varying channels has become an ubiquitous paradigm in the field of wireless communications. The necessity for this approach stems from the variable nature of the wireless medium as well as the variable nature of the incoming traffic to be transmitted over it. The adaptive resource allocation can alleviate the adverse influence of the variable channel and incoming traffic, and optimize the use of limited transmitter resources. Several wireless standards such as EDGE, IS-856, 802.11a,b and g, WCDMA and 1xEVDo already employ some sort of transmission adaptation based on the channel condition. The parameters that can be adapted at the transmitter are numerous and we will here name some of them. For example, it is possible to vary the transmission power which leads to *power control* algorithms. Some other options are adaptation of employed transmission rates, code rates or a combination of the above parameters.

In this paper we address the problem of finding optimal transmission adaptation policies for a single user data stream, their structure and related optimal costs. A very general approach to this problem is posed without restricting to the specific problem with a specified fixed costs. This approach is adopted in order to demonstrate that certain properties of adaptive transmission policies are applicable for different applications. We illustrate our general results with the examples well suited for transmission over wireless channels. Nevertheless, the results of this paper are quite general and can be applied to any adaptive transmission problems where resource-based costs have to be minimized while satisfying the constraint on certain QoS parameters.

The following model-related assumptions are employed in this paper. Firstly, it is assumed that processing is slotted and that scheduling decisions are made on regular intervals of length T. This assumption is commonly employed in practical wireless systems. By restricting ourselves to slotted processing, it is also possible to use the framework of constrained Markov decision processes and avoid the use of computationally demanding Semi-Markov decision processes. Secondly, the channel is assumed to be blockwise constant channel. In the context of a fading channel this implies the *block fading* channel model used by many authors (cf. [1]). The third model assumption is that the channel and source states are perfectly observable (precluding the use of Partially Observable Markov decision processes).

In the next subsection we will review some relevant related work on scheduling algorithms as well as general results on the properties of dynamic programming algorithms. Next, we will discuss the importance of the results presented in this paper as the foundation for devising efficient adaptive learning control algorithms for adaptive transmission problems. In Section II we pose the problem of choice of optimal adaptive transmission policies as a constrained Markov Decision Process (MDP) and define all the ingredients that constitute it. Some general mathematical results to be used in later sections are given in Section III. Our results on the monotonic increasing structure of the optimal deterministic and randomized policies are given in Section IV.

Notation: Upper case bold letters denote random variables, while lower case letters are reserved for the instances of random variables. Let $\mathbf{X}(y)$ denote the random variable $\mathbf{X}$ conditioned on the outcome y of the random variable $\mathbf{Y}$.

A. Related Work

Several recent publications deal with the problems of resource allocation adaptation for transmission over time-varying fading channels under constraints on the transmission delay [2], [3]. All of these results state the problem of finding the optimal policies as MDP and use standard dynamic programming tools to design optimal policies. Namely [2] addresses the problem of buffer-aware power and rate adaptation under delay constraints and demonstrate that adjoined multiobjective optimization problem can be posed as an unconstrained MDP. The structure of optimal deterministic rate adaptation policies for non-correlated channels have been analyzed in [4].

The authors are with the Department of Electrical and Computer Engineering, 2356 Main Mall, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada, ddjonin@ece.ubc.ca vikramk@ece.ubc.ca

0-7803-9568-9/05/$20.00 ©2005 IEEE

In order to establish our structural results on the optimal policies and costs we use general results on Constrained MDP's from the monograph [5]. We also use a result by Smith and McCardle [6] that gives a comprehensive survey of properties of the value function that are preserved with the iterations of the dynamic programming.

B. Paper Insights and Contributions

The main result of this paper is the formulation of a rate adaptive transmission scheduling problem using constrained Markov decision process framework. We give two examples on how this generic model can incorporate some known transmission adaptation problems. Based on this generic model and using supermodularity and convexity properties we establish several structural results on the optimal costs and policies.

For an active constraint on the buffer cost, it is shown in Theorem 2 that if the Lagrangian cost function of this model satisfies certain supermodularity and convexity properties, then the optimal deterministic scheduling policies are monotonically non-decreasing in the buffer occupancy. This simply means that irrespective of the channel and source states and their statistical description, the optimal scheduler takes more packets from the buffer with the increase of the buffer occupancy. This has practical implications for deriving efficient policy search algorithms (such as policy iteration [7]) as the search for the good policy can be constrained only to the subset of non-decreasing policies in the buffer size.

Furthermore, we derive the expression for the number of such deterministic optimal structured policies. For a general constraint we demonstrate that optimal randomized policies are probabilistic mixture of two deterministic monotone non-decreasing policies and present a simple algorithm that can produce these policies starting from deterministic policies. For large state spaces, the standard dynamic programming algorithms are hard to implement and finding of optimal scheduling policies is a very computationally demanding task. Therefore the significance of these structural results lies in the fact that we can restrict the search for a good suboptimal policy within only a certain constrained subset of policies.

II. Statement of the Problem

We state here the most general communication model that will be analyzed in this paper. The simpler analyzed examples can be obtained by choosing particular values for the parameters of this general model. We will use the framework of Constrained Markov Decision Processes (CMDP) to dynamically control our scheduling system that evolves in a stochastic manner. It is assumed that the scheduling decisions are made at the beginning of each block of duration T.

As in [5], a CMDP is defined with the tuple $\{\mathcal{S}, \mathcal{A}, P, c, d\}$. We now proceed to explain in detail the meaning of each component that constitutes the CMDP formulation of our communication model.

Let the set of states be denoted with $\mathcal{S}$. Each state $s \in \mathcal{S}$ is composed of three components $s = [h, b, f]$ and $\mathcal{S} = \mathcal{H} \times \mathcal{B} \times \mathcal{F}$. The first component of the state space $h \in \mathcal{H}$ is the current perfectly observable channel state. It is assumed that the channel is independent of the action, buffer state and incoming traffic. Therefore h is an uncontrollable component of the system state. It is assumed that channel states form an ergodic first order Markov chain with transition probabilities $p^h(h|h')$. Let $\mathbf{H}(h')$ be the random variable of the channel state in the next block as conditioned on the previous state h'.

Example 1: The channel can be modelled as finite state Markov Chain (cf. [8]). In this model it is assumed that transmission is performed in blocks of length T during which the channel state is represented by the current channel state supplied by the estimator at the receiver. The number of channel states is considered finite and the memory of the channel is limited to one. This model can be further simplified (as in [8]), only to allow the transitions between adjacent channel states.

Example 2: The complex channel gain $\tilde{h}_n$ at a certain transmission block n is modelled as p-th order autoregressive model

$$\tilde{h}_n = \sum_{j=1}^{p} \alpha_j \tilde{h}_{n-j} + v_n \tag{1}$$

where v_n is white (complex) Gaussian noise with variance σ^2 and mean μ. In this model, the state of the channel at n-th block is the vector $h_n = \{\tilde{h}_{n-1}, \ldots, \tilde{h}_{n-p}\}$. The coefficients α_j can be designed to approximate a specific fading autocorrelation function using Yule-Walker equations. If $p = 1$ this model represents an extension of the Markov Chain to continuous state variables. Assuming coherent detection at the receiver, the amplitude channel gain of the complex gain $\tilde{h}$ is $|\tilde{h}|$.

The second component of the state space is the current buffer state $b \in \mathcal{B}$ where set $\mathcal{B} = \{0, 1, , \ldots, L-1\}$ corresponds to all possible states of the buffer occupancy in packets.

The third component of the state space is $f \in \mathcal{F}$ which is the current perfectly observable incoming traffic state. Let the set $\mathcal{F} = \{0, 1, \ldots\}$ and let F be the maximum number of packets that can be received during a certain block with probability greater than some small ϵ. The incoming traffic state is assumed to be independent of the actions and the channel state and is therefore also uncontrollable. Incoming traffic state forms an ergodic Markov Chain with transition probabilities $p^f(f|f')$.

The set of actions $\mathcal{A} = \{1, \ldots, U\}$ comprises of all actions available in all the states. Let $\mathcal{A}_s \subset \mathcal{A}$ be the set of the actions that are available in the state s. Each action is interpreted as a choice of specific transmission rate/modulation order and/or power level at the transmitter. Let $\Psi(a)$ return the number of packets to be taken from the buffer provided that action a is applied. It is assumed that this function is increasing in a. An action a is available in state $s = [h, b, f]$ i.e.

$$a \in \mathcal{A}_s \Leftrightarrow \Psi(a) \leq b \text{ and } \Psi(a) \geq v(b - (L - 1 - F))$$

where $v(x)$ is a ramp function defined as $v(x) = x$ is $x \geq 0$ and $v(x) = 0$ otherwise. Note that the first condition prohibits sending more than available packets from the buffer while the second condition constraints the probability of overflows and packet dropping to ϵ. In general, the optimal actions in a state s of a CMDP are randomized and are given as random variable $\mathbf{A}$ with the support on $\mathcal{A}_s$.

The next component of the CMDP description is the set of stochastic transition matrices P that are defined for all actions $a \in \mathcal{A}$. The transition probability between the states $s = [h, b, f]$ and $s = [h', b', f']$ are given with

$$p(s'|s, a) = \delta(b' - \min(b + f - \psi(a), L))p^h(h'|h)p^f(f'|f) \tag{2}$$

where $a \in A_s$ and $\delta(x) = 1$ if $x = 0$ while $\delta(x) = 0$ otherwise. These transition probabilities form a set of stochastic matrices parameterized by the action a.

The cost $c(s, a)$ is the immediate cost of taking action a in the state s. It is assumed that this cost is independent of the next state of the system. The immediate cost can be given different interpretations and forms and we give the following example.

Example 3: Consider the channel as explained in Example 1. Let the actions $a \in \mathcal{A} = \{0, \ldots, \tilde{A}\}$ correspond to different 2^a-PSK modulation order used at the transmitter. It follows simply that $\Psi(a) = a$. For a fixed transmission power we can define the cost as average BER cost i.e.

$$c([b, h, f], a) = \int_{\gamma \in \Gamma_h} BER(\gamma, a)p^h(\gamma)d\gamma \tag{3}$$

where expectation is over signal to noise ratio (SNR) γ conditioned on the channel state being in state h i.e. $\gamma \in \Gamma_h$. The probability distribution function $p^h(\gamma)$ of the SNR conditioned on the channel state being h is considered known. Note that instantaneous bit-error rate $BER(\gamma, a)$ is commonly a convex for smaller values of γ (cf. [9]) and that can carry over to convexity and non-increasing property of $c(s[b, h, f], a)$ in h.

Example 4: Consider the channel as explained in Example 2 and assume that the transmitter is performing power adaptation of transmitted code words. Let the actions $a \in \mathcal{A} = \{0, \ldots, \tilde{A}\}$ correspond to rate a/N codes employed at the transmitter, where N is the number of symbols in the code word. It follows simply that $\Psi(a) = a$. Each rate action chosen at the transmitter determines specific transmission power for a fixed BER. We can define the transmission cost as the power

$$c([b, h, f], a) = \frac{\sigma^2}{\Gamma(BER)h}\left(2^{\Psi(a)/N} - 1\right), \tag{4}$$

that is necessary to achieve rate $\Psi(a)/N$ for specified bit error rate BER, and noise variance σ^2. The previous power cost comes from the expression for the rate R (cf. [2])

$$R = \log_2\left(1 + \Gamma(BER)\frac{P}{\sigma^2}\right), \tag{5}$$

that can be achieved for signal to noise ratio $\frac{P}{\sigma^2}$ with bit error rate BER. Here the SNR gap $\Gamma(BER)$ of a practical modulation corresponding to the classical Shannon capacity formula can be found in e.g. [10].

The immediate cost $d(s, a)$ is the optimization constraint and will be related to the buffer cost. As an example we can consider that this cost corresponds to the delay incurred by storing b packets in the buffer. Therefore the immediate buffer cost can be defined as

$$d([b, h, f], a) = \frac{b}{\bar{F}} \tag{6}$$

where $\bar{F}$ is the average number of incoming packets. Note that according to the Little's formula expectation of this constrained cost over the evolution of the system will give the average delay incurred in the buffer.

We will employ the average cost criteria [7] as a criteria for finding the optimal control of formulated CMDP. Let $\pi = \{\mathbf{A}_1, \mathbf{A}_2, \ldots\}$ be a sequence of randomized actions that constitutes a policy. Then we want to find the policy π that minimizes

$$C(\pi) = \mathbb{E}\left[\lim_{N \to \infty} \sup \sum_{i=1}^{N} \frac{1}{N} c(\mathbf{S}_n, \mathbf{A}_n)\right] \tag{7}$$

subject to buffer cost constraint

$$D(\pi) = \mathbb{E}\left[\lim_{N \to \infty} \sup \mathbb{E} \frac{1}{N} \sum_{i=1}^{N} d(\mathbf{S}_n, \mathbf{A}_n)\right] \leq \tilde{D}. \tag{8}$$

We will call the constraint (8) *active* if the equality holds in (8) for the optimal policy π. The expectation in (7) and (8) is with respect to system state $\mathbf{S}_n$, randomized actions $\mathbf{A}_n$ given the initial state distribution. Denote the optimal cost

$$C^* = \min_{\pi} C(\pi), \tag{9}$$

and any policy π that minimizes $C(\pi)$ be called the optimal policy.

We now review Theorem 12.7 from [5] to demonstrate that the optimal average cost and policy of the constrained Markov Decision Process can be found using an unconstrained MDP's and Lagrangian approach.

Theorem 1: The optimal cost function of the CMPD problem satisfies

$$C^* = \min_{\pi} \sup_{\lambda \geq 0} J(\pi, \lambda) - \lambda \tilde{D} = \sup_{\lambda \geq 0} \min_{\pi} J(\pi, \lambda) - \lambda \tilde{D} \tag{10}$$

where

$$J(\pi, \lambda) = \mathbb{E}\left[\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} c(\mathbf{S}_n, \mathbf{A}_n; \lambda)\right] \tag{11}$$

$$c(s, a; \lambda) = c(s, a) + \lambda d(s, a). \tag{12}$$

Note that the minimization in the rightmost expression in (10) can be done only over the set of deterministic policies. This important Theorem establishes that the CMDP problem can be solved by solving the appropriate Lagrangian unconstrained MDP problem and the relative value iteration algorithm available for this case.

III. Mathematical Preliminaries

Let us fix the value of the Lagrange multiplier to λ which has the meaning of fixing a certain value of delay constraint $\tilde{D}$. The relative value iteration algorithm (cf.[7]) can be defined as follows:

1. Select $V^0(s;\lambda)$, reference state s^* and specify ϵ.

2. For each $s \in S$, compute

$$V^{m+1}(s;\lambda) = \min_a \left[c(s,a;\lambda) + \sum_{s' \in S} p(s'|s,a) V^m(s';\lambda) \right] \tag{13}$$

3. Normalize $V(s;\lambda)$ for each $s \in S$ as

$$V^{m+1}(s;\lambda) = V^{m+1}(s;\lambda) - V^{m+1}(s^*;\lambda) \tag{14}$$

4. If

$$|V^{m+1}(s;\lambda) - V^m(s;\lambda)| < \epsilon \tag{15}$$

go to step 5. Otherwise increment m by 1 and return to step 2.

5. For each s choose the policy according to

$$\pi_\lambda(s) \in \arg\min_a \left[c(s,a;\lambda) + \sum_{s' \in S} p(s'|s,a) V(s';\lambda) \right] \tag{16}$$

and stop.

Let $V(s) = \lim_{m \to \infty} V^m(s)$. For a feasible action $a \in A_s$ in state s we can define the state-action value function

$$Q(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s) V(s'). \tag{17}$$

Function $Q(s,a)$ can be perceived as the equivalent immediate cost for the dynamic stochastic problem that can be used to find the optimal action in a given state. Now, the optimal policy can be found according to

$$\pi(s) = \arg\min_a Q(s,a) \tag{18}$$

while the optimal cost is

$$C^* = \mathbb{E}[V(\mathbf{S})]. \tag{19}$$

where the expectation is over the initial state distribution. The value iteration algorithm can also be used to determine the optimal randomized policy for a general constraint $\tilde{D}$. It has been shown in [11] that an optimal randomized policy for a discounted CMDP problem with a single constraint is *mixed policy* that is a probabilistic mixture of two stationary deterministic policies taken with probabilities q and $1-q$ where q depends on the constraint $\tilde{D}$. The same result has been later extended for average cost problems in [12].

A. Proof Guidelines

The proof dealing with establishing monotonic structure of scheduling policies in this paper follow the next four steps:

1) *Problem Formulation* The scheduling problem is formulated as an MDP with average cost criteria.

2) *Existence Results* For the stated problem it is shown that relative value iteration converges to the optimal relative value function.

3) *Supermodularity* It is shown that supermodular property of the value-action function is preserved with the iterations of the relative value iteration algorithm.

4) *Monotonicity of Policies* As the value-action function is supermodular it implies that the policy is monotonically increasing in a certain component of the state space.

We take a brief detour to give the explanations and definitions that will be necessary to formalize the proofs. Under the assumption that in each state s the support of $\mathbf{F}$ is including the set $\{0, \dots, \max \Psi(a)\}$ the step 2) follows from the observation that our models are *unichain* Markov Decision Processes which implies the convergence of the relative value iteration for bounded costs [7] .

Regarding the step 3) we need the definition of the submodular function.

Definition 1: A function $f : A \times X \times P \to \mathcal{R}$ is supermodular (has increasing differences) in (a,x) for a fixed parameter $p \in P$, if for all $a' \geq a$ and $x' \geq x$,

$$f(a',x';p) - f(a,x';p) \geq f(a',x;p) - f(a,x;p). \tag{20}$$

A function is $f : A \times X \times P \to \mathcal{R}$ is submodular (has decreasing differences) in (x,a) if the conditions of previous definition are satisfied and the inequality in (20) is flipped.

A central question for the interest of establishing the monotonic structure of the scheduling policies of step 4 is to identify when

$$\pi(x) = \arg\min_{a \in A} f(a,x;p) \tag{21}$$

will be non-decreasing in x for any parameter $p \in P$. This result was due to Topkis [13] and shows that submodularity of the function f in the pair (x,a) implies that $\pi(x)$ is non-decreasing function.

IV. Structured Policy Results

In this section we prove under quite general conditions that the optimal policy is monotonically increasing in the buffer state. In the course of this section we will assume the following: (1) the number of packets sent from the buffer after taking the action a is equal to the order of the action a i.e.$\Psi(a) = a$, (2) the scheduling decisions are performed periodically with period T, (3) CMDP is assumed to be unichain, [1] (4) relative value iteration converges. In the following theorem we first consider the case that the delay constraint $\tilde{D}$ is chosen such that a deterministic policy exists that is optimal for the given problem and satisfies the constraint, i.e. the constraint is active. Later this condition is relaxed and a similar property is shown for general constraints and randomized policies.

Theorem 2: For a certain $\lambda > 0$, let the immediate cost function $c([h,b,f],a;\lambda)$ given with (12) be submodular and jointly convex function of (b,a) and increasing function of b. Then the optimal policy $\pi([h,b,f])$ is non-decreasing function of b for $b < L - (F-1)$.

Proof:

[1] The unichain assumption is not necessary for our proofs if discounted costs [7] are used in lieu of average costs.

In order to prove that $\pi([h,b,f])$ is increasing function in b, we have to demonstrate that $Q([h,b,f],a;\lambda)$ is submodular function in the pair (b,a). We first notice that according to the statement of the theorem $c([h,b,f],a;\lambda)$ is submodular function of (b,a). Therefore we are only left to show that the second term of (17)

$$\begin{aligned} & Q_1([h,b,f],a;\lambda) && (22)\\ = \; & \sum_{h'\in\mathcal{H}}\sum_{f'\in\mathcal{F}} p^h(h'|h)p^f(f'|f)V([h',b-a+f',f'];\lambda) \end{aligned}$$

is submodular function of (b,a) for any h and f. The $\min(\cdot,L)$ operator in the previous equation is dropped since we are concerned only with the behavior of $Q([h,b,f],a;\lambda)$ when $b < L-(F-1)$ (cf.(2)).

Here we used the formulation of our model to simplify the equation (17). We first state the following Lemma whose proof is postponed for Appendix.

Lemma 1: $V([h,b,f];\lambda)$ is convex function of b for any h, f and λ given a jointly convex immediate cost function $c([h,b,f],a;\lambda)$ in (b,a).
Now, if function $V([h,b,f];\lambda)$ is convex in b, it can be shown that $V([h,b-a+f,f];\lambda)$ is also submodular in (b,a) for any h and f. This follows by noting that for a convex function $V([h,b,f];\lambda)$ of b it holds that

$$\begin{aligned} & V([h,x,f];\lambda)+V([h,y,f];\lambda) \geq \\ & V([h,\alpha x+(1-\alpha)y,f];\lambda)+V([h,(1-\alpha)x+\alpha y,f];\lambda) \end{aligned}$$

which is a direct consequence of the definition of convex function $V([h,b,f];\lambda)$. Substituting $x = b-a'+f, y = b'-a+f$ and $\alpha=(a-a')/(a-a'+b-b')$ in the previous equation and rearranging the terms we can get the following

$$\begin{aligned} & V([h,b'-a'+f,f];\lambda)+V([h,b'-a+f,f];\lambda) \geq \\ & V([h,b-a'+f,f];\lambda)+V([h,b-a+f,f];\lambda) \end{aligned}$$

For $a' \geq a$ and $b' \geq b$ this relation establishes the submodularity of $V([h',b-a+\bar{A}];\lambda)$ in the pair (b,a) for some channel and traffic states h and f. Furthermore, positive weighted sum of submodular functions is also submodular, which establishes the submodularity of $Q([h,b,f],a;\lambda)$ in (b,a) and concludes the proof. □

Remark 1: Note that in our transmission scheduling framework $c([h,b,f],a;\lambda)$ is submodular and jointly convex in (b,a). This is the consequence of $c([h,b,f],a;\lambda)$ being a sum of two convex functions that are dependent only on b and a respectively.

The results of the previous Theorem can significantly reduce the state of possible optimal policies in the given framework. Let us assume that there are A actions, L buffer states, K channel states and $F=1$ incoming traffic states (i.e. we assume that the incoming traffic is i.i.d.). Then the number of possible policies that possess the monotonically increasing structure $N(A,L,K)$ is equal to

$$N(A,L,K)=\left(A+\sum_{l=1}^{A-1}\binom{L-1}{l}\binom{A}{l+1}\right)^K$$

where the sum goes through all possible numbers of thresholds where the policy is increasing. This expression can be further simplified as

$$N(A,L,K)=\left(\frac{L+A-1!}{(A-1)!L!}\right)^K \qquad (23)$$

The reduction in the size of the structured policy space is quite significant as compared to the set of all possible policies. Consider for example $L=100, A=5, K=2$ for which the number of non-structured policies A^{LK} is on the order of 10^{140} while the number of structured policies is on the order of 10^{13}. This reduction in the number of possible policies we intend to explore further when the channel and/or incoming traffic model is unknown and neurodynamic programming with adaptive learning is employed to obtain the optimal policy.

Theorem 1 demonstrates that the optimal deterministic policy for a CMDP can be found using the relative value iteration (RVI) in case that the constraint is active. However, we can still pose the question of finding the suitable Lagrangian multiplier λ that satisfies a certain constraint with equality. Note that the average constraint for the optimal policy π^*_λ for Lagrangian multiplier λ can be given with

$$D(\pi^*_\lambda;\lambda)=\mathbb{E}\lim_{N\to\infty}\sup\left[\frac{1}{N}\sum_{i=1}^{N}d(\mathbf{S}_n,\mathbf{A}_n)\right]. \qquad (24)$$

In order to get the average constraint $D(\pi^*_\lambda;\lambda)$ for a specified λ we have to apply the relative value iteration. As was shown in [2] when the Lagrangian multiplier λ trade-offs between the cost and the constraint, $D(\pi^*_\lambda;\lambda)$ is step-wise continuous decreasing function of λ. An example of this dependance is shown in Figure 1 given the transmission costs of Example 4 and constrained buffer costs of (6).

A simple algorithm designed to find the smallest λ (that will be called λ^*) such that the constraint (8) is satisfied can be formulated as following

$$\lambda_{n+1}=\lambda_n-\epsilon_n(D(\pi^*_\lambda;\lambda)-\tilde{D}) \qquad (25)$$

where the step $\epsilon_n=\frac{1}{n}$ and λ_1 is sufficiently large number. The convergence to λ^* is ensured as the function

$$\int_0^\lambda\left(D(\pi^*_\lambda;\lambda)-\tilde{D}\right)d\lambda \qquad (26)$$

is piece-wise linear concave function that attains its global maximum at λ^* and its derivative is $D(\pi^*_\lambda;\lambda)$. Therefore the algorithm (25) is just a gradient descent algorithm.

We now explore the structure of scheduling policies for a general constraint $\tilde{D}$ with randomized policies. Let $v_P(s,a)$ be a probability that an action a is taken in a state s under a certain randomized policy P. Under the conditions of the previous Theorem and using the result [11] it is easy to show that the optimal randomized policy is a mixture of two deterministic monotonically increasing policies with $v^*([b_1,h,f],a) \geq v^*([b_2,h,f],a)$ for any $b_1 \geq b_2$. Furthermore, optimal randomized policy can have at most two atoms of mass in the distribution of $v^*([b,h,f],a)$ for any fixed b,h and f.

At the end of this section we demonstrate how to employ the algorithm (25) and estimated parameter λ^* to find the optimal randomized policy for any feasible constraint $\tilde{D}$ with RVI. Let $a_\pi^-(s) = \min_a\{a : v_\pi(s,a) > 0\}$ and $a_\pi^+(s) = \max_a\{a : v_\pi(s,a) > 0\}$. Then it is easy to show that $a_\pi^-([b_1,h,f]) \geq a_\pi^-([b_2,h,f])$ and $a_\pi^+([b_1,h,f]) \geq a_\pi^+([b_2,h,f])$ for any $b_1 \geq b_2$. Note that, according to [14], the number states with randomized policies in a unichain MDP model with only one constraint is no more than 1. In view of [11], we perturb the parameter λ^* by some $\delta\lambda$ to get $\lambda^- = \lambda^* - \delta\lambda$ and $\lambda^+ = \lambda^* + \delta\lambda$. Next we find the optimal deterministic policies $\pi^*_{\lambda^-}$ and $\pi^*_{\lambda^+}$ and their respective average constrained costs $D^- = D(\pi^*_{\lambda^-})$ and $D^+ = D(\pi^*_{\lambda^+})$. The optimal randomized policy is to be a mixture of two deterministic policies and let parameter q determine the probability of taking the policy $\pi^*_{\lambda^-}$ and $1-q$ be the probability of taking the policy $\pi^*_{\lambda^+}$. Now, parameter q can be estimated such that $qD^- + (1-q)D^+ = \tilde{D}$. Note that this optimal randomized policy can be implemented as round-robin policy as described in [14].

V. Conclusions

This paper establishes general structural results of optimal deterministic and randomized policies for the constrained MDP formulation of the transmission adaptation problems. Given the particular fading channel state, it is shown that the optimal policy allocates to transmit more packets with the increase of the transmit buffer occupancy.

A particularly interesting and useful extension of the work presented in this paper is to devise efficient adaptive control algorithms that can adaptively improve the control policies in unknown environments. Since the state space of the MDP for our transmission adaptation model can be huge, it is of interest to employ the structure of the optimal policies in order to speed up the convergence of algorithms such as the Q-learning or TD learning [15].

Appendix

Proof of Lemma 1

The proof follows by mathematical induction and using the relative value iteration. Choose $V^0([h,b,f];\lambda)$ that is a convex function of b. Now, we will show that convexity of $V^m([h,b,f];\lambda)$ implies the convexity of $V^{m+1}([h,b,f];\lambda)$ in b. According to the value iteration algorithm

$$V^{m+1}([h,b,f];\lambda) = \min_a \Bigg[c([h,b,f],a;\lambda) + \sum_{h' \in \mathcal{H}} \sum_{f' \in \mathcal{F}} p^h(h'|h) p^f(f'|f) V^m([h', b-a+f', f'];\lambda) \Bigg]. \tag{27}$$

Based on convexity of $V^m([h,b,f];\lambda)$, it can be easily shown that $V([h', b-a+f', f'];\lambda)$ is jointly convex in (b,a). Furthermore, since $c([h,b,f],a;\lambda)$ is jointly convex in (b,a) the sum of convex functions before minimization in the previous equation is also jointly convex. Using the property that $g(x) = \min_a f(x,a)$ is a convex function of x for a jointly convex function f of (x,a), it follows that $V^{m+1}([h,b,f];\lambda)$ is also convex function of $m+1$. □

References

[1] E. B. J. Proakis and S. S. (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. on Inform. Theory*, pp. 2619 – 2692, Oct. 1998.

[2] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. on Inform. Theory*, pp. 1135 – 1149, May 2002.

[3] D. Rajan, A.Subharwal and B.Aazhang, "Delay and rate constrained transmission policies over wireless channels," *in Proc. of Globecom Conf.*, pp. 806 – 810, 2001.

[4] M. Goyal, A. Kumar and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," *in Proc. of INFOCOM*, pp. 311 – 320, 2003.

[5] E. Altman, *Constrained Markov Decision Processes: Stochastic Modeling*. London: Chapman and Hall CRC, 1999.

[6] J. E. Smith and K. F.McCardle, "Structural properties of stochastic dynamic programs," *Operations Research*, pp. 796 – 809, Sep.-Oct. 2002.

[7] M. L. Putterman, *Markov Decision Procsses: Discrete Stochastic Dynammic Programming*. New York: John Wiley & Sons, 1994.

[8] H. S. Wang and N. Moayeri, "Finite-state Markov channel- a useful model for radio communication channels," *IEEE Trans. on Vehicular Tech.*, pp. 163 – 171, Feb. 1995.

[9] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*. New York: John Wiley & Sons, 2000.

[10] J. M. Cioffi, "A multicarrier premier," *available at http://www.stanford.edu/group/cioffi/pdf/multicarrier.pdf*, Nov. 1999.

[11] F. J. Beutler and K. W. Ross, "Optimal Policies for Controlled Markov Chains with a Constraint," *Journal of Math. Anal. and Applications*, vol. 112, pp. 236 – 252, 1985.

[12] L.I.Sennott, "Constrained average cost Markov decision chains," *Probability in the Engineering and Information Sciences*, vol. 7, pp. 69 – 83, 1993.

[13] D. M. Topkis, *Supermodularity and Complementarity*. Prineceton University Press, Prineceton, NJ, 1998.

[14] K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints," *Operations Research*, vol. 37, pp. 474 – 477, 1987.

[15] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, Massachusetts: Athena Scientific, 1996.

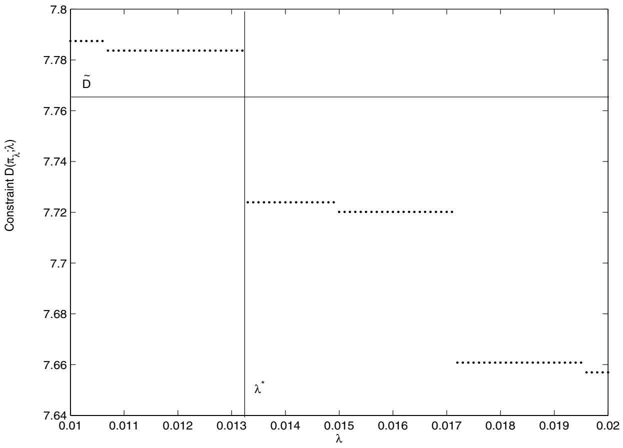

Fig. 1. The plot of constrained cost $D(\pi^*_\lambda;\lambda)$ associated with the optimal policy π^*_λ vs. Lagrange multiplier λ.