An SMDP Approach to Optimal PHY Configuration in Wireless Networks

Mark Shifrin, Daniel S. Menasché, Asaf Cohen, Omer Gurewitz, Dennis Goeckel
Ben Gurion University of the Negev, Federal Univ. of Rio de Janeiro, Univ. of Massachusetts at Amherst

***Abstract*—In this work, we study the optimal configuration of the physical layer in wireless networks by means of Semi-Markov Decision Process (SMDP) modeling. In particular, assume the physical layer is characterized by a set of potential operating points, with each point corresponding to a rate and reliability pair; for example, these pairs might be obtained through a now-standard diversity-vs-multiplexing tradeoff characterization. Given the current network state (e.g., buffer occupancies), a Decision Maker (DM) needs to dynamically decide which operating point to use. The SMDP problem formulation allows us to choose from these pairs an optimal selection, which is expressed by a decision rule as a function of the number of awaiting packets in the source's finite queue, channel state, size of the packet to be transmitted. We derive a general solution which covers various model configurations, including packet size distributions and varying channels. For the specific case of exponential transmission time, we analytically prove the optimal policy has a threshold structure. Numerical results validate this finding, as well as depict muti-threshold policies for time varying channels such as the Gilber-Elliot channel.**

I. Introduction

Recent advances in coding and modulation have allowed communication systems to approach the Shannon limit on a number of communication channels; that is, given the channel state (e.g. the signal-to-noise ratio of an additive white Gaussian noise - AWGN - channel), communication near the highest rate theoretically possible while maintaining low error probability is achievable. However, in many communication systems, particularly wireless communication systems, the channel conditions which a given transmission will experience are unknown. For example, consider the case of slow multipath fading without channel state information at the transmitter [1]. Because of the uncertainty in the level of multipath fading, it is possible that the rate employed at the transmitter cannot be supported by the channel conditions, hence resulting in packet loss or "outage". The outage capacity [1], which gives the rate for various outage probabilities, captures the tradeoff in such a situation. If a low rate is employed, it is likely that the channel conditions will be such that transmission at that rate can be supported (low outage); if a higher rate is employed, the probability is higher that the channel conditions will be such that transmission at that rate cannot be supported (high outage). Below we will indicate the wide range of physical layers that can be addressed through such an approach and how this characterization can be systematically derived. Given this characterization, a crucial question is at which point to operate the physical layer given information available about the current state of the network.

Modern wireless networks are extremely dynamic, with channel parameters and traffic patterns changing frequently. Consequently, the key question addressed here is how can a sender *dynamically decide* on the best physical layer strategy, given the channel and traffic parameters available to it, as well as its own status. For example, consider a sender required to decide among the aforementioned physical layer strategies: an approach incurring high packet loss yet a small transmission time, or one possibly having a lower packet loss but a larger transmission time. In this paper, we derive a framework for a Decision Maker (DM) wishing to maximize the system throughput by choosing the appropriate physical layer setting, while taking into account as many system parameters as possible, in this case, delay, packet losses and its current packet backlog status. The DM faces a choice of achieving increased success probability provided additional transmission time, and one would expect that this decision will be made with the queue status in mind, as a full queue causes new arrivals to be rejected, incurring potential throughput loss. Thus, our goal is to rigorously analyze this tension, and identify the optimal strategy.

The tradeoff between rate and reliability is a fundamental characteristic of the physical layer, and we are interested in this formulation largely because it captures much more than the simple point-to-point single-antenna communication used for illustration in the first paragraph above. To provide a systematic method to consider how this characterization might be derived, consider the the now-standard diversity-multiplexing tradeoff approach originally applied to point-to-point multi-antenna systems [2] but now extensively extended beyond that. In particular, the diversity-multiplexing approach captures the fundamental tradeoff between rate (multiplexing) and reliability (diversity) for a number of interesting physical layer choices, including: (1) point-to-point multiple-input and/or multiple-output (MIMO) systems [2], where the transmitter can decide whether to send multiple streams ("multiplex") from the multiple antennas or to send a single stream with redundancy ("diversity"), or a combination of the two; (2) half-duplex relay channels (e.g. [3]), where the transmitter can decide to use the relay, or not, and how to allocate time to the transmit and receive functions of the half-duplex relay. We are particularly interested in this latter example, and we will adopt terminology from a classical problem in relaying [4] to help clarify the competing options in succeeding sections. However, we hasten to remind the reader that the results apply much more generally to the diversity-multiplexing protocol for any physical layer.

Our model is also relevant for sources transmitting packets of variable sizes. In particular, once the size of the leading packet in the buffer is known prior to the the transmission, selecting the point of the discussed tradeoff also selects the transmission time, which has a crucial impact on the future buffer occupancy. If the buffer is full and arriving packets are immediately rejected, the potential throughput associated with these packets is lost. Clearly, the decision rule and the tradeoff structure are not straightforward in such cases. Nonetheless, we consider them in our model.

A. Main contribution

Our main contributions are as follows.

Problem formulation: We formulate the problem of optimal dynamic PHY configuration for the transmitter with long-lived packets influx by a SMDP. The model is presented in generality, capturing finite buffer size, variable packet size, variable channel state, general transmission time distribution and a decision space which is associated with possible PHY configuration.

Value function derivation: We derive the equations for the value function of the SMDP in several interesting cases. These equations are obtained in a tractable form, allowing a solution by simple value iteration.

Threshold policy characterization: When transmission times are exponentially distributed, we prove there exists an *optimal policy with a threshold structure*; that is, the source should make use of the more reliable option if the number of pending packets is lower than a given threshold, and transmit with the higher rate option otherwise. We also show that the value function in this case is concave and increasing.

Numerical investigation: We explore different scenarios by simulations. In particular, we validate the threshold policy and concavity for the exponential case and observe similarity to this structure in other cases as well. In the case of a variable channel (e.g., the Gilbert-Elliot channel) we observe a *multi-threshold* policy which is described by having a separate threshold for each channel mode.

While a preliminary paper [5] gave basic numerical results for a related model, to the best of our knowledge, this work is the first to *analyze* this problem for a finite buffer at the source and generality of all other system parameters.

B. Related work

The pioneering works on the relay channel date back to the seventies [6], when T. Cover and E. Gamal posed the problem of determining the capacity region of the full-duplex relay channel. Since then, there has been progress in determining the capacity region for the degraded case, as well as for several MIMO settings [7] or a class of erasure channels [8]. Cooperative strategies and their performance in relay networks were considered in [9]. Nonetheless, in its full generality the capacity region of the relay channel is still unkown.

Yeh and Berry [10] considered control policies which account for queue dynamics in order to optimize both scheduling and routing. The throughput optimization consists of maximizing the stability region of a set of networked queues by means of maximum differential backlog algorithms. Our approach is different, as we consider the control of a finite queue, hence inherently stable.

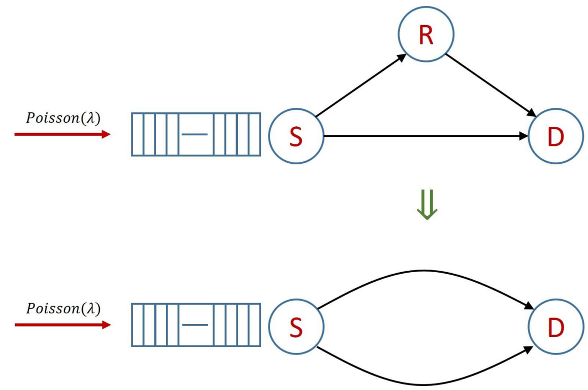

Fig. 1. Relay channel logical model

Recently, Urgaonkar and Neely [11] considered a constrained resource allocation problem in a relay network under stringent delay constraints. The reliability versus delay trade-off in finite buffered networks was addressed in [12] and in wireless networks in [13]. Our work differs from such previous studies as we consider a different approach to model the relay channel, namely semi-Markov decision processes (SMDP), which allows us to analyze a broader set of settings and channels as opposed to those considered in [11], [12], [13].

For a basic introduction to SMDPs we refer the reader to [14]. A number of previous works on SMDPs have focused on establishing the existence of optimal policies of threshold type under a variety of settings [15], [16], [17], [18]. To the best of our knowledge, none of these works have considered the optimality of threshold policies for SMDP models of the relay channel.

The general trade-off between PHY rate and reliability has various important applications apart from the already mentioned basic relay channel. For example, the trade-off between multiplexing and diversity was discussed in [19] in the context of MIMO channels, in [20] in the context of multiple-access channel with fading, and in [21] in the context of cognitive radio sensing techniques.

Space diversity can be achieved by transmitting simultaneously through two channels, minimizing the effects of fading in a single slot, while time diversity can be achieved since fading also varies over time, hence different schemes can be used at different times. Works such as Berry and Gallager [22] and Collins and Cruz [23] accounted for time diversity and delay constraints, while works such as Scaglione, Goeckel and Laneman [24] accounted for space diversity. The SMDP framework introduced in this paper accounts for both space and time diversity.

II. System model and metrics

We assume the application level at the source (S) produces packets to be transmitted to the destination (D). The structure of packets is discussed in sequel. The source attempts to communicate the packet at the head of the queue, formed in a *finite sized buffer*, to the destination by a choice among the possible operating points of the physical layer. In the case there is an ongoing transmission of an earlier produced packet, the new packets are fed into the queue, space permitting. The packets arriving at full buffer are rejected and the retransmission details are taken care of by the application level. For presentation and analytical simplicity, we assume a pair of operating points at all decision epochs. While this choice applies cleanly to the simple half-duplex relay channel example, the extension to the situation of more than two possible operating points, as might be appropriate for other PHY layer architectures, is straightforward. Also note that in the case of continuous intervals of operating points, the model can be addressed by known techniques for continuous action space, e.g., discretization. The tension is clear: a higher rate choice that uses less resources (time) but at a higher packet loss probability, or a lower rate choice that exploits more resources to obtain a lower packet loss probability.

Transmission modes: We assume two possible operating points, and we use a nomenclature motivated by the application of the diversity-multiplexing tradeoff to a classic half-duplex relay channel choice: the more reliable and lower relaying rate and outage probability from a prototypical early cooperative diversity protocol, and the less reliable and higher rate corresponding to the direct choice. In all scenarios mentioned in this paper, we will always denote the more reliable path with a lower rate as path "a", while the less reliable path with a higher rate will be denoted as path "b". Denote the rates R_a and R_b, where $R_a < R_b$. The corresponding packet losses are denoted as p_a and p_b, where $p_a < p_b$ at all times.

We consider packets of random length, hence, the rates of the two options yield transmission time probability density functions $g^b(t)$ and $g^a(t)$, which are readily calculated from the packet length distribution, R_a and R_b. Note that in the case of a half duplex relay model $g^a(t)$ refers to the entire path associated with choice "a", even if the latter is subdivided into two separate paths. Therefore, we aim to analyze a simplified configuration as Figure 1 demonstrates. Note that the upper illustration particularly fits the relay channel problem, while the lower one can also refer to the general tension between two PHY settings associated with two different propagation paths; the upper path corresponds to the choice of reliability (diversity) while the lower path corresponds to the choice of rate (multiplexing).

Variable channel: We assume that a variable channel can be modeled by a Markov process; that is, the *current channel statistics* are independent of the past given the last state of the channel. These dynamics reflect well-known channel models (e.g., i.i.d. fading, Gilbert-Elliot model). Consider a sequence of fading values observed across source packets, each value corresponding to a channel (or, separately, to each path) state. We assume that the statistics are known to the DM and are constant for a period which is significantly longer than the longest possible packet transmission time. Hence the DM can obtain the packet losses associated with each such fading value for each potential PHY configuration. Note that we allow generality of these probabilities by assuming they can be different for different transmission modes. Clearly, $\sum_{h'} p^u(h'|h) = 1$. Also, for the empty buffer state, we allow these probabilities to be distinct from those at other buffer states, hence capturing the time it takes until the first packet arrival to the queue. While these assumptions are approximations, they conform to the well-known realistic slow fading model, or quasi-static channels. Denote the channel state by h, and the channel transition probability from state h to state h' by $p^u(h'|h)$, where the superscript u, $u \in \{a, b\}$ denotes the transmission mode.

Arrival process: We assume that the *chunks* arrival process at the source is a Poisson process with intensity λ. The size of the buffer is known to the DM and is given by B chunks. The packet is ready to be transmitted once the newly arrived chunk indicates that it is the last one in the packet. Therefore, this model captures the following packet sizing options:

1) Each packet can have a size of one or more chunks. After each transmission, the corresponding number of chunks is subtracted from the buffer.
2) The packets of all sizes occupy space of exactly one chunk. (For example, packet descriptors are stored in the buffer, or the chunks are large enough to fit the maximal possible packet size.) After each transmission, exactly one chunk is subtracted from the buffer.

We allow *two modes* of packet size impact. In the first mode, the packet size is unknown prior to the transmission, and the transmission times are merely distributed with g^u defined above. Note that this can model the scenario where application produces chunks and starts the transmission before the last chunk of the packet has been produced. In the second mode, the size can be sampled prior to the transmission. Then, the size of the packet to be transmitted is a part of the *state* and has impact on the decision. Denote the size of the packet in this case by k. Then, *packet size transition probabilities*, denoted by $q^u(k'|k)$, stand for the probability of having packet of size k' at the head of the line after transmitting packet of size k by acting u, where $u \in \{a, b\}$. And, as in the case of the channel transition probabilities, these probabilities can be different for the empty buffer state to allow for the possibility of a different size distribution for the first packet to arrive into the empty buffer. We assume a finite set of possible packet sizes such that $\sum_{k'} q^u(k'|k) = 1$ for each possible action u. Hence, in the first mode, g^u can be merely assumed as deterministic given the packet sizes, or be independent of packet size and selected from other system parameters. Note that q^u can capture complex packet arrival patterns. For example, a sequence of big packets which is likely to be followed by sequence of small patterns and vice versa can be modeled provided the appropriate values for q^u are selected. We assume that the packet arrival process and variable channel process are independent.

General performance criterion : We now define the performance criterion, which will suit our SMDP formulation as defined in the next section. Denote the ending time of m-th transmission attempt as σ_m, $m \in \{0, 1, \cdots\}$. We assume that the rewards are only added up at transmission ending times. Hence, in the case the transmission at time σ_m was successful, the reward worth of r_m was positive. Otherwise, r_m was equal to zero. Define average reward over infinite horizon

$$J^A = \lim_{N \to \infty} \frac{1}{N} \mathbb{E} \sum_{m=1}^{N} r_m$$

variable	description
λ	arrival rate
B	buffer capacity (maximum number of packets in the system, including the packet being transmitted)
s	SMDP state (buffer state and channel state) Note: as state transitions occur after departures, the buffer state after a transition ranges between 0 and $B-1$
$V(s)$	value function at state s, $V(s) = \max\{V^{(r)}(s), V^{(d)}(s)\}$
$V^{(u)}(s)$	cost associated with decision u at state s, $u \in \{a, b\}$
$\beta_{u,s}$	discount associated with action u
k	number of frames per packet
$\tau_{u,s}$	mean time to transmit via channel u at state s
$\mu_{u,s}$	transmission rate of channel u, $\tau_{u,s} = 1/\mu_{u,s}$
p_u	packet loss probability associated with channel u, $u \in \{a, b\}$
$g^{\pi}(t)$	pdf of transmission time then using policy π
$P^{u}(j\|i,t)$	probability of $j-i$ arrivals after t time units, given a buffer initially filled with i packets and action u is taken
$p^{u}(h'\|h)$	channel transition probabilities, given action u is taken
$q^{u}(k'\|k)$	packet size transition probabilities, given action u is taken

TABLE I
SMDP NOTATION.

Thus, our goal is the following:

Find the *dynamic physical layer setting selection policy*, which takes as input the current occupancy of the buffer and information (if any) on the current state of the channel and the packet size at the source's queue head, and provide as an output a decision on which transmission path to employ, such that J^A is maximized.

We formulate the described problem by a Semi Markov Decision Process (SMDP), as formally defined in the next section.

III. SMDP-BASED FORMULATION AND SOLUTION

Denote SMDP with average cost functional by tuple $\{\mathbf{S}, \mathbf{A}, \mathbf{P}, r\}$, where the components stand for the state space, the action space, the transition probabilities and the reward function. We consider the most general case where the state $s \in \mathbf{S}$ is expressed by the triplet $(n, h, k) \in \mathbb{R}^3$ where n stands for the number of information chunks in the system, h stands for the medium (channels) state and k stands for the size of the packet (measured in number of chunks) which is to be transmitted in the upcoming transmission. The state transitions happen after transmission over either possible path ("a" or "b") ends, or, in the case there was no ready packet for the transmission (e.g. the buffer was empty), the chunk which just has arrived indicates that the only packet awaiting now in the queue is ready. We assume that a *transmission mode* (i.e. transmission decision) with corresponding parameters (e.i. corresponding transmission rates) to each transmission path is attributed. Right after the state transition the decision which sets the transmission mode of the current packet is made, and its transmission is immediately started.

The probability of having j chunks after transmission interval of length t is denoted by $\varrho(j|i,t)$ and is governed by a Poisson distribution with mean λt, e.g. if $j < B+1, i > 0$ then $j = i - k + m$, where m is the number of arrivals during time t. The action space is defined by $\mathbf{A} = \{a, b, 0\}$, standing for transmission mode "a", transmission mode "b", and abstaining from transmission. In the case the buffer was empty or action 0 was selected, no transmission is initiated, while the next decision will be performed right after the next packet is ready. Otherwise, the actions are taken right after the accomplished transmissions. The instantaneous gain at the end of a successful transmission is given by $r = k$, in the case the reward is set according to the packet size, and $r = 1$ in the case successfully transmitted packets of all sizes have the same value.

In order to simplify the forthcoming analysis, we use a discounted infinite horizon SMDP defined by tuple $\{\mathbf{S}, \mathbf{A}, \mathbf{P}, r, \gamma\}$, where the first four components are exactly the same as in the average cost formulation and γ is a discount factor. The discounted cost is given by the following

$$J^{\pi}(s_0) = \mathbb{E}\sum_{m=0}^{\infty} e^{-\gamma\sigma_m} r_m = \mathbb{E}r^{\pi}(s_0) + \mathbb{E}\sum_{m=1}^{\infty} e^{-\gamma\sigma_m} r_m = = \mathbb{E}r^{\pi}(s_0) + J^{\pi,1}(s_0) \quad (1)$$

The connection between discounted criterion and average infinite reward criterion is well understood. In particular, under mild conditions for γ, they possess the *similar optimal policy*, which is named the Blackwell optimal policy. See [25] for the details. Note that the second term of (1) is a result of applying dynamic programming and it is given by $\mathbb{E}_{(\sigma_1,s_1)}[e^{-\gamma\sigma_1}V(s_1)]$. The superscript π stands for the policy and the value function is given by $V(s_0) = \max_{\pi} J^{\pi}(s_0)$, for all $s_0 \in \mathbf{S}$. We now expand for all possible transitions. Denote the transition probability $P(s_1|s_0, \pi(s_0))$. Accounting for number of arrivals, channel transitions and leading packet size transitions, for $s_0 = (h, k, i)$, write

$$P(s_1|s_0, \pi(s_0)) = P((h', k', j)|(h, k, i), \pi(s_0)) = = q^{\pi(s_0)}(k'|k) p^{\pi(s_0)}(h'|h) \varrho(j|i,t) \quad (2)$$

The Bellman equation is given by (1), setting $J^{\pi,1}(s_0)$ as follows

$$J^{\pi,1}(s_0) = \quad (3)$$

$$= \sum_{s_1=(h',k',j)} V(s_1) \int_0^{\infty} e^{-\gamma t} P(s_1|s_0, \pi(s_0)) g^{\pi(s_0)}(t) dt \quad (4)$$

$$= \sum_{h'} \sum_{k'} \sum_{j=i-k}^{B-k+1} V(s_1) \int_0^{\infty} e^{-\gamma t} P(s_1|s_0, \pi(s_0)) g^{\pi(s_0)}(t) dt. \quad (5)$$

The first (i.e the outer) summation in (5) is over all possible next channel states. It is degenerated if the channel is fixed. The second summation is over all possible packet sizes to be transmitted at state s_1. If packet size is unknown, or all packet sizes are equal, this sum degenerates. The third summation is over the number of arrivals to the queue during the first transmission. The integration accounts for the average time the first transmission interval takes. Hence, the transmission time pdf $g^{\pi(s_0)}$ may depend on the action taken in state s_0. The expected reward at state s_0 is added up at the *end of the transmission* and is calculated by $\mathbb{E}r^{\pi}(s_0) = \int r^{\pi(s_0)} e^{-\gamma t} g^{\pi(s_0)}(t) dt$. Note that $r^0 = 0$, while $r^a = k(1-p_a)$, $r^b = k(1-p_b)$. Also note that in the case $i < k$, (i.e. the first and the only packet in the buffer is still waiting for remaining chunks to arrive) the $g^{\pi(s_0)}(t)$ is given by distribution time of remaining $k-i$ packets to arrive and the reward is zero. Observe that each product of probabilities of the form $q^{\pi(s_0)}(k'|k) p^{\pi(s_0)}(h'|h) \varrho(j|i,t)$ is a weight of the corresponding value function at the next state s_1. The optimal value function is retrieved by the maximization over all admissible policies $V(s_0) = \max_\pi J^\pi(s_0)$.

We bring next three examples, for which the corresponding Bellman equations can be written in the tractable form and the value functions can be explicitly calculated, e.g. by value iteration [14]. Note that for all examples the boundary conditions at buffer limits are separately written.

A. Exponential transmission times

Next, we consider exponentially distributed transmission times. Let $1/\mu_u$ be the expected transmission time through channel u, $u \in \{a, b\}$, as estimated by the controller.

The state space is one-dimensional. The system state is characterized by the number of packets at the source, and is denoted by s, $s \in \{0, 1, \ldots, B-1\}$. We assume all packets are equally valued with associated reward 1. Letting $g^u(t) = \mu_u e^{-\mu_u t}$ in (4), where superscript u stands for the action taken, $u \in \{a, b\}$, yields

$$J^{u,1}(j) = V(B-1) \int_0^\infty e^{-\gamma\tau_u} \varrho(B|j,t) \mu_u e^{-\mu_u \tau_u} d\tau_u + \\ + \sum_{i=j-1}^{B-1} V(i) \left(\int_0^\infty e^{-\gamma\tau_u} \frac{e^{-\lambda\tau_u}(\lambda\tau_u)^{i-j+1}}{(i-j+1)!} \mu_u e^{-\mu_u\tau_u} d\tau_u \right)$$

where $\varrho(B|j,t) = \left(1 - \sum_{i=j-1}^{B-1} \frac{e^{-\lambda\tau_u}(\lambda\tau_u)^{i-j+1}}{(i-j+1)!}\right)$. $\varrho(B|j,t)$ denotes the probability that, during a service, at least one packet is lost due to buffer overflow given that there are j packets in the system prior to the beginning of the service. Note that $\int_0^\infty t^n e^{-st} dt = \frac{n!}{s^{n+1}}$. Thus, for $0 < j \le B-1$ we have,

$$J^u(j) = C(1-p_u)\frac{\mu_u}{\mu_u+\gamma} + \sum_{i=j-1}^{B-1} V(i) \frac{\lambda^{i-j+1}\mu_u}{(\gamma+\mu_u+\lambda)^{i-j+2}} + \\ + V(B-1)\left(\frac{\mu_u}{\gamma+\mu_u} - \sum_{i=j-1}^{B-1} \frac{\lambda^{i-j+1}\mu_u}{(\gamma+\mu_u+\lambda)^{i-j+2}}\right) \tag{6}$$

The optimal value function is given by $V(n) = \max_u J^u(n)$, $n = 1, \ldots, B-1$. Note that the value function at the boundary condition $n = B-1$ is obtained directly from the equations above, whereas $V(0)$ is given by

$$V(0) = \frac{\lambda}{\lambda+\gamma} V(1). \tag{7}$$

Note that due to the summation in (6) the value function of the SMDP at each state may depend on all other states. For this reason, a direct analysis of the SMDP is cumbersome. To circumvent this challenge, we rely on an MDP formulation equivalent to the SMDP presented above. The resulting Bellman equations are simple, hence analysis and identification of optimal policies of threshold type is plausible.

1) MDP formulation for the exponential case: Next, we define the states of the MDP and their corresponding value functions. The transition diagram of the MDP is illustrated in Figure 2. The definition of the state space is inspired by the MDP admission control example presented in [14, chpt. 11].

Our goal is to leverage the Markovian structure of the problem when the times between all events are exponentially distributed. To this aim, we modify the state space to $\{0, 1, \ldots, B-1\} \cup \{0, 1, \ldots, B-1\} \times \{a, b\}$. Whereas under the general SMDP framework transitions occurred after every departure, we now consider transitions that occur *after every departure or arrival*. States (n), $n = 0, \ldots, B-1$, are achieved after a departure (service completion), whereas states (n, u), $n = 0, \ldots, B-1$, $u \in \{a, b\}$ are achieved after arrivals.

The system transitions to state (n) after a service completion that leaves behind n packets in the buffer. As soon as the system reaches state (n), $n = 1, \ldots, B-1$, the controller decides between transmitting the head-of-line packet through channels a or b. The mean residence time at state (n), $n = 1, \ldots, B-1$ is $1/(\mu_a + \lambda)$ or $1/(\mu_b + \lambda)$, if actions a or b are chosen, respectively. If a new packet arrives and encounters an idle system, the system transitions from state (0) either to state $(0, a)$ or $(0, b)$, depending on whether action a or b is chosen. The mean residence time at state (0) is $1/\lambda$.

Next, we consider a new packet that arrives to encounter a busy system. We assume that the arrival finds a total of n packets in the buffer (accounting for the packet being transmitted), and a packet being transmitted through channel u, $u \in \{a, b\}$. Immediately after the arrival the system transitions to state (n, u). At state (n, u) the controller does not take any actions, as the only possibility is to continue the ongoing transmission. The mean residence time at state (n, u) is $1/(\mu_u + \lambda)$.

Let $\mathcal{P}(s_1|s_0, u_0)$ be the transition probability from state s_0 to state s_1, given that action u_0 is taken. Then, the system dynamics is captured by $\mathcal{P}(s_1|s_0, u_0)$ as follows.

$\mathcal{P}(s_1|s_0, u_0) =$

$$= \begin{cases} 1, & s_0 = (0), s_1 = (0, u_0) & \text{(8a)} \\ \dfrac{\lambda}{\lambda+\mu_u}, & s_0 = (n), s_1 = (n, u_0), u = u_0 \text{ or} & \text{(8b)} \\ & s_0 = (n-1, u), s_1 = (n, u) & \text{(8c)} \\ \dfrac{\mu_u}{\lambda+\mu_u}, & s_0 = (n), s_1 = (n-1), u = u_0 \text{ or} & \text{(8d)} \\ & s_0 = (n, u), s_1 = (n) & \text{(8e)} \\ 0, & \text{otherwise} & \text{(8f)} \end{cases}$$

where $n = 1, \ldots, B-1$. Note that (8) depends on u_0 only through transitions (8a), (8b) and (8d). At states (n, u), transitions (8c) and (8e) are independent of u_0, which means that the controller does not need to take any actions at those states. Note that the at states (n, a) and (n, b) the variable n does not account for the arriving packet, which will be admitted to the system in case the buffer is not full.

Instantaneous rewards are accumulated once transmissions are finished. Equivalently, such rewards are added at the beginning of a transmission, multiplied by the corresponding expected discount. In what follows, we let c_a and c_b be the expected instantaneous reward received when actions a and b are taken, respectively.

2) MDP Bellman equations for the exponential case: Next, we introduce the value functions and Bellman equations which characterize the solution of the MDP model. Recall that state

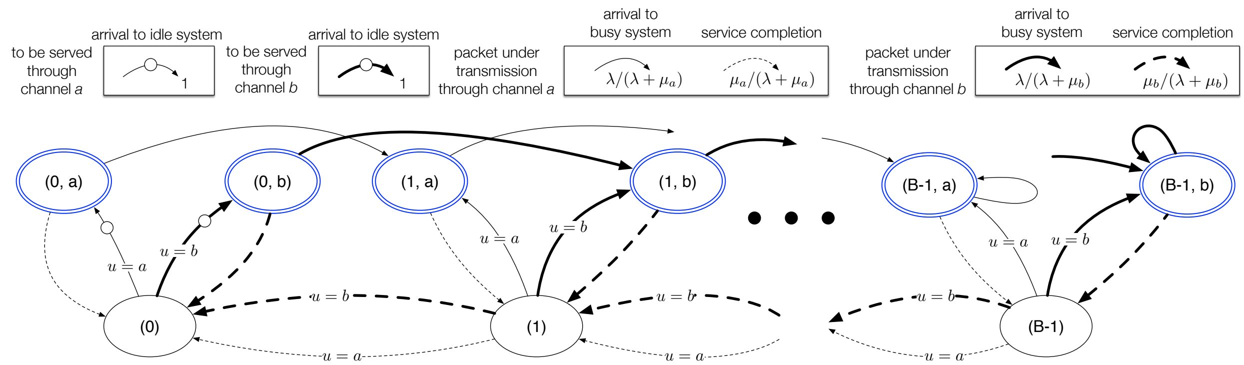

Fig. 2. MDP model of a system with buffer capacity B. Expected instantaneous reward c_a (resp., c_b) is received when action a (resp., b) is taken.

(n) is achieved right *after a departure* (transmission completion) which leaves n packets at the buffer, $n \in \{0, 1, 2, \dots, B-1\}$. The corresponding value function is denoted by V_n. Recall also that state (n, u) is achieved right *after an arrival* to a system with n packets, $n \in \{0, 1, 2, \dots, B-1\}$, including the one which is currently being transmitted on path u, $u \in \{a, b\}$. If $n = 0$, the transmission of the arriving packet is immediately started through u. The value function corresponding to state (n, u) is denoted by $V_n^{u,A}$ ("A" stands for "arrival").

We are mainly interested in the values of states (n), i.e., our main goal is to obtain V_n, $n \in \{0, 1, 2, \dots, B-1\}$, both because the decisions are made only in these states and because they are *comparable with the SMDP values*.

Let $\delta_a = (\mu_a + \lambda + \gamma)^{-1}$, $\delta_b = (\mu_b + \lambda + \gamma)^{-1}$, $\beta_b = (\mu_b + \gamma)^{-1}$,$\beta_a = (\mu_a + \gamma)^{-1}$and $\bar{\delta} = (\gamma + \lambda)^{-1}$. We write next Bellman equations for the value function for states at arrival events. These equations define the corresponding operators $\mathbb{A}_a$ and $\mathbb{A}_b$, which act in the space of function from $\{0, \cdots, B\}$ to $\mathbb{R}$.

$$
\begin{aligned}
V_n^{b,A} &= \mu_b \delta_b V_n + \lambda \delta_b V_{n+1}^{b,A} = \mathbb{A}_d V_n^{b,A} \qquad (9)\\
V_n^{a,A} &= \mu_a \delta_a V_n + \lambda \delta_a V_{n+1}^{a,A} = \mathbb{A}_r V_n^{a,A},
\end{aligned}
$$

Denote by $\mathbb{T}, \mathbb{T}_a, \mathbb{T}_b$ the operators acting on the space of functions from $\{0, \cdots, B\}$ to $\mathbb{R}$. When applied to state n, these operators yield the following equations:

$$
\begin{aligned}
V_n^{b,D} &= \mathbb{T}_b V_n = \mu_b \delta_b V_{n-1} + \lambda \delta_b V_n^{b,A} + (1 - p_b)\beta_b\\
V_n^{a,D} &= \mathbb{T}_a V_n = \mu_a \delta_a V_{n-1} + \lambda \delta_a V_n^{a,A} + (1 - p_a)\beta_a \qquad (10)
\end{aligned}
$$

The maximization is performed by $V_n = \max(V_n^{b,D}, V_n^{a,D}) = \mathbb{T} V_n = \max(\mathbb{T}_b V_n, \mathbb{T}_a V_n)$. At the buffer limit boundary B we have

$$
\begin{aligned}
V_B^{b,A} &= \mu_b \delta_b V_{B-1} + \lambda \delta_b V_B^{b,A}\\
V_B^{a,A} &= \mu_a \delta_a V_{B-1} + \lambda \delta_a V_B^{a,A}, \qquad (11)
\end{aligned}
$$

and at the empty buffer,

$$
V_0^{b,D} = V_0^{a,D} = \max(\lambda \bar{\delta} V_0^{b,A}, \lambda \bar{\delta} V_0^{a,A}). \qquad (12)
$$

Arrivals that find an empty buffer are subject to the effect of the controllers' currrent decision,

$$
V_0^{b,A} = V_0^{a,A} =
$$
$$
\max(\mu_b \delta_b V_0^{b,D} + \lambda \delta_b V_1^{b,A} + c_b, \mu_a \delta_a V_0^{a,D} + \lambda \delta_a V_1^{a,A} + c_a)
$$

Note that $V_0^{b,A} = V_0^{a,A}$ holds because at state (0) no packet is currently being transmitted. The derivation of the equations above are presented in the on-line version of this paper [26]. See that each operator application updates one state by using values of two other states. This can be followed from the arrows in the diagram of Figure 2.

variable	description
c_u	expected instantaneous reward associated to decision u.
(n, u)	state right after arrival, when current active transmission is through u and n packets are found by the arrival.
(n)	state following a transmission completion, when n packets are left in the buffer. A decisions is made at this state.
$V_n^{u,A}$	value function at state (n, u).
V_n	value function at state (n), $V_n = \max\{V_n^{a,D}, V_n^{b,D}\}$.
$V_n^{u,D}$	state-action value function for action u at state (n).
$\mathbb{A}_u$	operator applied over arrivals (acts on $V^{u,A}$).
$\mathbb{T}_u$	operator applied over departures (acts on $V^{u,D}$).
$\mathbb{T}$	operator that maximizes the outcome of $\mathbb{T}_u$ over u.

TABLE II
MDP NOTATION ($u \in \{a, b\}$, $n \in \{0, \dots, B-1\}$).

In Section IV we will use the MDP formulation to identify the threshold type structure of the optimal policy.

B. Deterministic transmission times and known packet sizes

Consider a source which samples the packet size before the transmission. Consider two possible sizes, denoted by k_1 and k_2, both taking in the buffer exactly one slot, i.e. case 2) in section II. We also assume equal rewards for both sizes. Then, the state is given by $s = \{n, k\}$, $n \in \{0, \cdots, B\}$ and $k \in \{k_1, k_2\}$. We assume that packet size dynamics is given by a discrete Markov Chain with transition probabilities $q(k'|k)$. For simplicity we also assume $\mathbf{A} = \{a, b\}$, that is, the packet, if ready, has to be transmitted. Denote the deterministic transmission time of packet of size k as $\tau^{\pi(n,k)}$. The packet loss probabilities are p_a and p_b. Denote $s_0 = (i, k)$, $s_1 = (j, k')$. Then,

$$
J^{\pi,1}(i, k) = \sum_{k'} \sum_{j=i-1}^{B} e^{-\gamma \tau^{\pi(s_0)}} Q((s_1)|s_0, \pi(s_0)) V(j, k') \qquad (13)
$$

where $Q((s_1)|s_0, \pi(s_0)) = q^{\pi(s_0)}(k', k)\varrho(j|i, \tau^{\pi(s_0)})$, and $\mathbb{E} r^{\pi(i,k)} = (1 - p_{\pi(i,k)}) e^{-\gamma \tau^{\pi(n,k)}}$. Hence, the value functions are $V(i, k_m) = \max_\pi \{J^{\pi(i,k_m)}\}$, $m \in \{1, 2\}$. The boundary condition, then the buffer is empty

$$
V(0, k) = \mathbb{E} e^{-\gamma t_\lambda} V(1, k) = \sum_{k'} \frac{\lambda}{\gamma + \lambda} \big(q_0(k'|k, 0) V(1, k)\big)
$$

C. Gilbert-Elliot relay channel with uniform transmission times

Assume the packet sizes cannot be sampled, but are known to have a uniform distribution over all channels. Consider two states of the *entire medium*, denoted by A and B. Namely, both relay and direct routes are fully dependent and can simultaneously be in one of the states. That is, the channel state refers to the entire medium. The channel state is sampled prior to each upcoming transmission, and is modeled as part of the state space, which determines the packet losses in current transmission slot. Namely, $s = \{n, h\}$, $n \in \{0, \cdots, B\}$ and $h \in \{A, B\}$. Denote $u = \pi(n, h)$. We assume that channel dynamics can be expressed by discrete Markov Chain, which corresponds to the Gilbert–Elliott (GE) channel. The packet transmission time $\tau^{a,h}$ (resp. $\tau^{b,h}$) over relay (resp. direct) channel is uniformly distributed. The uniform distributions intervals, which depend on the channel state and the action, are given by $[\alpha_{u,h}, \beta_{u,h}]$. For simplicity, we assume that in the buffer all packets occupy exactly one slot and are equally rewarded. We assume that channel transition probabilities, denoted by $p^u(h'|h)$, depend on the action $u \in \{r, d\}$, where relay (resp. direct) transmission corresponds to the transmission mode "a" (resp. "b"). Denote $s_0 = (i, h),\ s_1 = (j, h)$. Then,

$$J^{u,1}(i,h) = \sum_{j=i-1}^{B} \sum_{h'} V(j,h') \int_{\alpha_{u,h}}^{\beta_{u,h}} e^{-\gamma t} Q(s_1|s_0, \pi(s_0)) dt \tag{14}$$

where $Q(s_1|s_0, \pi(s_0)) = \frac{1}{\beta_{u,h} - \alpha_{u,h}} \varrho(j|i,t) p^u(h'|h)$, and $\mathbb{E}r^u = (1 - p_u) \int_{\alpha_{u,h}}^{\beta_{u,h}} e^{-\gamma t} \frac{1}{\beta_{u,h} - \alpha_{u,h}} dt = (1 - p_u) \frac{e^{-\gamma(\beta_{u,h} - \alpha_{u,h})}}{\beta_{u,h} - \alpha_{u,h}}$. Note that the probability to have full buffer after end of transmission is given by $\varrho(B|i,t) = (1 - \sum_{i=j-1}^{B-1} \varrho(j|i,t))$. Finally, the value function for state s is given $V(s) = \max_u \{J^u(s)\}$.

IV. Structure of optimal policies

The structure of the optimal policy has particular importance, for several reasons. First, in order to assess the resources needed for the policy implementation at wireless nodes. Next, structural properties can be exploited by learning algorithms in order to significantly reduce the complexity of optimal policy search. This is especially useful for the system with large state-space. For example, once the policy is proven to possess a *threshold* structure in one of the dimensions of a state space, the data to hold for the policy consists of only single scalar. Moreover, the configuration of similar systems can be analytically or heuristically based on the existing one, e.g. by means of reinforcement learning aimed to policy improvement. We aim to identify threshold policies for the SMDP models and solutions presented above. For the exponential case, we analytically prove the threshold property. We finally compare by simulations the thresholds associated with other service time distributions. To this end, we state our main analytical result:

Theorem 1. *The problem with exponentially distributed transmission times modeled by MDP is solved by the optimal policy of a threshold type. Namely, there exists a unique threshold t, $0 \le t \le B$, such that the optimal policy is to transmit via path "a" for all states where $n \le t$ and to transmit via path "b" otherwise.*

By the equivalence of the value functions at departures in MDP and SMDP formulations trivially the following holds.

Corollary 1. *The exponential relay problem modeled in section III by SMDP is solved by the optimal policy of threshold type.*

To this end, let $\mathcal{S}$ be a set where each of its elements is a five-tuple of B-dimensional vectors denoted by $\{U, U^{b,A}, U^{a,A}, U^{b,D}, U^{a,D}\}$ satisfying the following properties

1) the difference $U_n^{a,D} - U_n^{b,D}$ is non-decreasing in n, $n \in \{0, \cdots, B\}$
2) $\{U^{r,d}, U^{d,d}, U^{b,A}, U^{a,A}\}$ are concave in $n \in \{1, \cdots, B\}$,
3) $\{U, U^{b,A}, U^{a,A}\}$ are non-decreasing in $n \in \{0, \cdots, B\}$,
4) $\{U, U^{b,A}, U^{a,A}\}$ have their slope bounded by some positive constant K, that is, $U_n - U_{n-1} < K$, $U_n^{a,A} - U_{n-1}^{a,A} < K$ and $U_n^{b,A} - U_{n-1}^{b,A} < K$ and, For the proof of the theorem we will need the following lemma.

Lemma 1. *The operators $\mathbb{A}_b$,$\mathbb{A}_a$,$\mathbb{T}$ preserve properties 1)-4).*

Due to the lack of space, the proof of the lemma appears in the online version of the paper [26].

Proof of Theorem 1. The proof is based on a well known result that operators associated with Bellman equation are contracting [14], that is, using the maximum metric $\parallel U \parallel = \max_x |U(x)|$ it holds $a \parallel U - W \parallel < \parallel \mathbb{T}U - \mathbb{T}W \parallel$ for some $0 < a < 1$. Hence, the operators defined above are contraction mappings, equipped with the metric $\rho(U;W) = ||U - W||$ in a complete metric space. Since $\mathcal{S}$ is a complete metric space and the operators are strict contractions, they have corresponding fixed points (e.g. in [27, Theorem V.18]). Now since $\mathcal{S}$ is not empty (one can easily construct such functions; the technical details are omitted), the functions which are in $\mathcal{S}$ and have the operators $\mathbb{A}_a, \mathbb{A}_b, \mathbb{T}$ applied on them, by lemma 1 stay in $\mathcal{S}$. By contraction, the repetitious application brings the result infinitesimally close to the fixed points of $\mathbb{A}_a, \mathbb{A}_b, \mathbb{T}$. Recall that the value functions $V_n, V_n^{b,A}, V_n^{a,A}$ are the unique solution of *all* functions, including those that in $\mathcal{S}$, acting from $n \in \{0, \cdots, B\}$ to $\mathbb{R}$, to the *same* equations; (trivially, the mild conditions for uniqueness and existence, see e.g. [14, Chapter 6.2], apply). As a result, $\{V, V^{b,A}, V^{a,A}, V^{b,D}, V^{a,D}\}$ coincide with these fixed points and they are in $\mathcal{S}$. In particular, $V^{b,D}$ and $V^{a,D}$ possess property 4), which is equivalent of having at most one policy switch state. This proves the proposition. □

V. Numerical results

In this section, we report numerical results on the shape of the value functions obtained through value iteration. Our goals are to 1) illustrate how different system parameters impact the performance of threshold policies and 2) to numerically investigate the optimality of multi-threshold optimal policies for the Gilbert-Elliot channel.

In Figure 3 we compare the value functions and threshold policies for channels associated to exponential, deterministic and uniform transmission times. The mean transmission rates were set to $\mu_a = 9$ and $\mu_b = 12$, under the low and high loads, respectively. The support of the uniformly distributed transmission times was set to $\alpha_u = 0.2\frac{1}{\mu_u}, \beta_u = 1.8\frac{1}{\mu_u}$. We considered both a low load ($\lambda = \mu_a = 9$) and a high load ($\lambda = \mu_b = 12$) regime. Vertical lines show the thresholds where the policy determines a switch from a to b. Observe that under high load the thresholds are significantly smaller than under low load. This is because under low loads it is key to avoid buffer underflows, which cause a reduction in system throughput.

For the high load, note at the states close to B the value function becomes nearly constant. This is explained by the fact that at all these states the average time until the buffer empties is large. Hence, at these states the penalty due to a potential non-realized future discounted reward is negligible.

Also note that the numerical results validate our formal results on the concavity of the value function for the exponential case. In addition, the value function for the two other cases are also concave, a result which is interesting on its own.

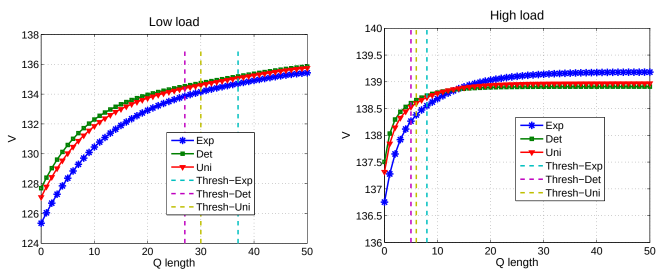

Fig. 3. Value functions (V) at different buffer states (Q). Average transmission rates and packet losses are given by $\mu_a = 9, p_a = 0.25$ and $\mu_b = 12, p_b = 0.42$. Right: high load ($\lambda = 13$), left: low load ($\lambda = 9$).

Figure 4 illustrates the multi-threshold policies obtained when solvig the SMPD model under deterministic and uniform transmission times with Gilbert-Elliot channels. The two channel types are denoted as A and B. Each transmission time distribution corresponds to *two value functions*, for channels A and B. Hence, each value function implies its own threshold. Observe that the value functions for A and B are very close to each other. Nonetheless, the thresholds can be easily distinguished. Under the Gilbert-Elliot channel model, for all the scneraios considered we were always able to find a separate threshold for each channel type. While a rigorous analysis of the multi-threshold policy is subject for future work, the numerical analysis presented here can be used to devise heuristics to be concurrently applied with value iteration, aiming towards faster convergence.

VI. Conclusion

We have proposed an SMDP model for optimal PHY configuration, derived equations for the value function for several interesting cases, and formally shown structural properties of the optimal policy when transmission times are exponentially distributed. In particular, we have shown the existence of optimal policies of threshold type. The numerical solution of the proposed model indicates the good performance of multi-threshold policies for Gilbert-Elliot channels. Showing the optimality of the latter under general settings is an interesting open problem.

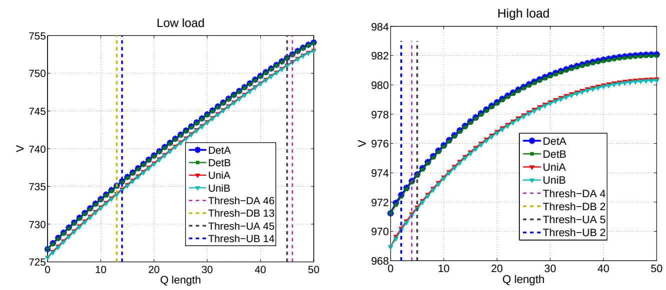

Fig. 4. Value functions (V) at different buffer states (Q): $\mu_a = 10, p_{a,A} = 0.2, p_{a,B} = 0.35$ and $\mu_b = 15, p_{b,A} = 0.3, p_{b,B} = 0.4$. High load ($\lambda = 15$) and low load ($\lambda = 10$), with threshold values inside the legend.

References

[1] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
[2] L. Zheng and D. N. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *Trans. Info. Theory*, 2003.
[3] S. Karmakar and M. K. Varanasi, "The diversity-multiplexing tradeoff of the dynamic decode-and-forward protocol on a mimo half-duplex relay channel," *Trans. Info. Theory*, vol. 57, no. 10, 2011.
[4] J. N. Laneman, D. N. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *Info. Theory, IEEE Trans. on*, vol. 50, no. 12, pp. 3062–3080, 2004.
[5] D. S. Menasché, D. Goeckel, and D. Towsley, "Optimal forwarding strategies for the relay channel under average delay constraints," in *ACM Wireless S3*, 2010, pp. 1–4.
[6] T. M. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *Trans. Info. Theory*, pp. 572–584, 1979.
[7] B. Wang, J. Zhang, and A. Host-Madsen, "On the capacity of mimo relay channels," *Trans. Info. Theory*, vol. 51, no. 1, 2005.
[8] R. Khalili and K. Salamatian, "On the achievability of cut-set bound for a class of erasure relay channels," in *ISITA*, 2004.
[9] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *Info. Theory, IEEE Trans. on*, 2005.
[10] E. M. Yeh and R. A. Berry, "Throughput optimal control of cooperative relay networks," *Info. Theory, Trans.*, 2007.
[11] R. Urgaonkar and M. J. Neely, "Delay-limited cooperative communication w/reliability constraints in wireless networks," *Trans. Info. Theory*, 2014.
[12] S. Supittayapornpong and M. J. Neely, "Achieving utility-delay-reliability tradeoff in stochastic network optimization with finite buffers," *arXiv preprint arXiv:1501.03457*, 2015.
[13] E. Altman, K. Avratchenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. S. Menasché, "Constrained stochastic games in wireless networks," in *Globecom*. IEEE, 2007, pp. 315–320.
[14] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.
[15] G. Koole, "Structural results for the control of queueing systems using event-based dynamic programming," *Queueing Systems*, 1998.
[16] B. Hajek, "Optimal control of two interacting service stations," *IEEE Transactions on Automatic Control*, vol. 29, no. 6, pp. 491–499, 1984.
[17] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*. John Wiley & Sons, 2009, vol. 504.
[18] J. Walrand, *An introduction to queueing networks*. Prentice Hall, 1988.
[19] R. Heath and A. Paulraj, "Diversity versus multiplexing in narrowband mimo channels," *submitted Dec*, 2002.
[20] D. N. C. Tse, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple-access channels," *Trans. Info. Theory*, 2004.
[21] Y.-C. Liang, K.-C. Chen, G. Y. Li, and P. Mahonen, "Cognitive radio networking and communications," *Trans. Vehic. Techn.*, 2011.
[22] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *Trans. Info. Theory*, 2002.
[23] B. E. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Allerton*, 1999.
[24] A. Scaglione, D. L. Goeckel, and J. N. Laneman, "Cooperative communications in mobile ad hoc networks," *Signal Proc. Mag.*, 2006.
[25] D. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995, vol. 2, no. 2.
[26] M. Shifrin, D. Menasche, A. Cohen, O. Gurewitz, and D. Goeckel, "An smdp approach to optimal phy configuration in wireless networks," *On-line version, Available on ArXiv:1601.06859*, 2016.
[27] M. Reed and B. Simon, *Methods of modern mathematical physics: Functional analysis*. Gulf Professional Publishing, 1980, vol. 1.