An Adaptive Opportunistic Routing Scheme for Wireless Ad-hoc Networks

A.A. Bhorkar, M. Naghshvar, T. Javidi, *Member, IEEE*, and B.D. Rao, *Fellow, IEEE*

***Abstract*—In this paper, a distributed adaptive opportunistic routing scheme for multi-hop wireless ad-hoc networks is proposed. The proposed scheme utilizes a reinforcement learning framework to opportunistically route the packets even in the absence of reliable knowledge about channel statistics and network model. This scheme is shown to be optimal with respect to an expected average per packet reward criterion.**

The proposed routing scheme jointly addresses the issues of learning and routing in an opportunistic context, where the network structure is characterized by the transmission success probabilities. In particular, this learning framework leads to a stochastic routing scheme which optimally "explores" and "exploits" the opportunities in the network.

***Index Terms*—Opportunistic routing, reward maximization, wireless ad-hoc networks.**

I. Introduction

Opportunistic routing for multi-hop wireless ad-hoc networks has seen recent research interest to overcome deficiencies of conventional routing [1]–[6] as applied in wireless setting. Motivated by classical routing solutions in the Internet, conventional routing attempts to find a fixed path along which the packets are forwarded [7]. Such fixed path schemes fail to take advantages of broadcast nature and opportunities provided by the wireless medium and result in unnecessary packet retransmissions. The opportunistic routing decisions, in contrast, are made in an online manner by choosing the next relay based on the actual transmission outcomes as well as a rank ordering of neighboring nodes. Opportunistic routing mitigates the impact of poor wireless links by exploiting the broadcast nature of wireless transmissions and the path diversity.

The authors in [1] and [6] provided a Markov decision theoretic formulation for opportunistic routing. In particular, it is shown that the optimal routing decision at any epoch is to select the next relay node based on an index summarizing the expected-cost-to-forward from that node to the destination. This index is shown to be computable in a distributed manner and with low complexity using the probabilistic description of wireless links. The study in [1], [6] provided a unifying framework for almost all versions of opportunistic routing such as SDF [2], Geographic Routing and Forwarding (GeRaF) [3], and EXOR [4]. The variations in [2]–[4] are due to the authors' choices of cost measures to optimize. For instance an optimal route in the context of EXOR [4] is computed so as to minimize the expected number of transmissions (ETX). GeRaF [3] uses the smallest geographical distance from the destination as a criterion for selecting the next-hop.

The opportunistic algorithms proposed in [1]–[6] depend on a precise probabilistic model of wireless connections and local topology of the network. In practical setting, however, these probabilistic models have to be "learned" and "maintained." In other words, a comprehensive study and evaluation of any opportunistic routing scheme requires an integrated approach to the issue of probability estimation. Authors in [8] provide a sensitivity analysis in which the performance of opportunistic routing algorithms are shown to be robust to small estimation errors. However, by and large, the question of learning/estimating channel statistics in conjunction with routing remains unexplored.

In this paper, we investigate the problem of opportunistically routing packets in a wireless multi-hop network when zero or erroneous knowledge of transmission success probabilities and network topology is available. Using a reinforcement learning framework, we propose an adaptive opportunistic routing algorithm which minimizes the expected average per packet cost for routing a packet from a source node to a destination. Our proposed reinforcement learning framework allows for a low complexity, low overhead, distributed asynchronous implementation. The most significant characteristics of the proposed solution are:

- It is oblivious to the initial knowledge of network.
- It is distributed; each node makes decisions based on its belief using the information obtained from its neighbors.
- It is asynchronous; at any time any subset of nodes can update their corresponding beliefs.

The idea of reinforcement learning has been previously investigated for conventional routing in Ad-hoc networks [9] [10]. In [9], a ticket-based probing scheme is proposed for path discovery in MANETs to reduce probe message overhead. This heuristic can be viewed as a very special case of our work where the probabilistic wireless link model is replaced with a deterministic link model. In [10], the authors attempts to find an optimal path dynamically in response to variations in congestion levels in various parts of the network. As discussed in the conclusion, the issue of congestion control remains open and entails further research.

The rest of the paper is organized as follows: In Section II, we discuss the system model and formulate the problem. Section III formally introduces our proposed adaptive routing algorithm, d-AdaptOR. We then state and prove the optimality theorem for d-AdaptOR algorithm in Section IV. In Section V,

Manuscript received June 27, 2009. This work was partially supported by the UC Discovery Grant #com07-10241, Ericsson, Intel Corp., QUALCOMM Inc., Texas Instruments Inc., CWC at UCSD, and NSF CAREER Award CNS-0533035.

The authors are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (email: abhorkar@ucsd.edu; naghshvar@ucsd.edu; tjavidi@ucsd.edu; brao@ucsd.edu).

we present the implementation details and practical issues. We perform simulation study of the proposed algorithm in Section VI. Finally, we conclude the paper and discuss future work in Section VII.

We end this section with a note on the notations used. On the probability space $(\Omega, \mathcal{F}, P)$, let $I : \Omega \to \{0, 1\}$ denote the indicator random variable (with respect to $\mathcal{F}$), such that for all $\omega \in \Omega$, $A \in \mathcal{F}$, $I(A) = 1$ for all $\omega \in A$, and $I(A) = 0$ for all $\omega \notin A$. For a vector $x \in \mathbb{R}^D$, $D \geq 1$, let $x(l)$ denote the l^{th} element of the vector. Let $||x||_v$ denote the weighted max-norm with positive weight vector v, i.e. $||x||_v = \max_l \frac{|x(l)|}{v(l)}$. Let $\mathbf{1} \in \mathbb{R}^D$ denote the vector with all components equal to 1. We also use the notation X^n to represent the first n random elements of the random sequence $\{X_k\}_{k=1}^{\infty}$.

II. System Model

We consider the problem of routing packets from a source node o to a destination node d in a wireless ad-hoc network of $d+1$ nodes denoted by the set $\Theta = \{o, 1, 2, \ldots, d\}$. The time is slotted and indexed by $n \geq 0$ (this assumption is not technically critical and is only assumed for ease of exposition). A packet indexed by $m \geq 1$ is generated at the source node o at time τ_s^m according to an arbitrary distribution with rate $\lambda > 0$.

We assume a fixed transmission cost $c_i > 0$ is incurred upon a transmission from node i. Transmission cost c_i can be considered to model the amount of energy used for transmission, the expected time to transmit a given packet, or the hop count when the cost is equal to unity.

Given a successful transmission from node i to the set of neighbor nodes S, the next (possibly randomized) routing decision includes 1) retransmission by node i, 2) relaying the packet by a node $j \in S$, or 3) dropping the packet all together. If node j is selected as a relay, then it transmits the packet at the next slot, while other nodes $k \neq j, k \in S$, expunge that packet.

We define the termination event for packet m to be the event that packet m is either received by the destination or is dropped by a relay before reaching the destination. We define termination time τ_e^m to be a random variable when packet m is terminated. We discriminate amongst the termination events as follows: We assume that upon the termination of a packet at the destination (successful delivery of a packet to the destination) a fixed and given positive reward R is obtained, while no reward is obtained if the packet is terminated (dropped) before it reaches the destination. Let r_m denote this random reward obtained at the termination time τ_e^m, i.e. it is either zero if the packet is dropped prior to reaching the destination node or R if the packet is received at the destination.

Let $i_{n,m}$ denote the index of the node which transmits packet m at time n. The routing scheme can be viewed as selecting a (random) sequence of nodes $\{i_{n,m}\}$ for relaying packets $m = 1, 2, \ldots$.[1] As such, the expected average per packet reward associated with routing packets along a sequence of $\{i_{n,m}\}$ upto time N is:

$$J_N = \mathbf{E}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}}\right\}\right], \qquad (1)$$

where M_N denotes the number of packets terminated upto time N and the expectation is taken over the events of transmission decisions, successful packet receptions, and packet generation times.

Problem (P) Choose a sequence of relay nodes $\{i_{n,m}\}$ in the absence of knowledge about the network topology such that J_N is maximized as $N \to \infty$.

In the next section we propose d-AdaptOR algorithm which solves Problem **(P)**. The nature of the algorithm allows nodes to make routing decisions in distributed, asynchronous, and adaptive manner.

Remark The problem of opportunistic routing for multiple source-destination pairs can be effectively decomposed to the problem above where routing from one node to a specific destination is addressed.

III. Distributed Algorithm

In this section we present the description of d-AdaptOR scheme. In the rest of the paper, we let $\mathcal{N}(i)$ to denote the set of neighbors of node i including node i itself. Let $\mathfrak{S}^i$ denote the set of potential reception outcomes due to a transmission from node $i \in \Theta$, i.e. $\mathfrak{S}^i = \{S : S \subseteq \mathcal{N}(i), i \in S\}$. We refer to $\mathfrak{S}^i$ as the state space for node i's transmission. Furthermore, let $\mathfrak{S} = \cup_{i\in\Theta}\mathfrak{S}^i$. Let $A(S)$ denote the space of all allowable actions available to node i upon successful reception at nodes in S, i.e. $A(S) = S \cup \{f\}$. Finally, for each node i we define a reward function on states $S \in \mathfrak{S}^i$ and potential decisions $a \in A(S)$ as

$$g(S, a) = \begin{cases} -c_a & \text{if } a \in S \\ R & \text{if } a = f \text{ and } d \in S \\ 0 & \text{if } a = f \text{ and } d \notin S \end{cases}.$$

A. Overview of d-AdaptOR

As discussed before, the routing decision at any given time is made based on the successful outcomes and involves retransmission, choosing the next relay, or termination. Our proposed scheme makes such decisions in a distributed manner via the following three-way handshake between node i and its neighbors $\mathcal{N}(i)$.

1) At time n node i transmits a packet.
2) Set of nodes S_n^i who have successfully received the packet from node i, transmit acknowledgment (ACK) packets to node i. In addition to the node's identity, the acknowledgment packet of node $k \in S_n^i$ includes a control *message* known as *estimated best score* (EBS) and denoted by Λ_{max}^k.
3) Node i announces node $j \in S_n^i$ as the next transmitter or announces the termination decision f in a forwarding (FO) packet.

The routing decision of node i at time n is based on an adaptive (stored) score vector $\Lambda_n(i, \cdot, \cdot)$. The score vector

[1] Packets are indexed according to the termination order.

$\Lambda_n(i,\cdot,\cdot)$ lies in space $\mathbb{R}^{v_i}$, where $v_i = \sum_{S\in\mathfrak{S}^i}|A(S)|$, and is updated by node i using the EBS messages Λ^k_{max} obtained from neighbors $k \in S^i_n$. Furthermore, node i uses a set of counting variables $\nu_n(i,S,a)$ and $N_n(i,S)$ and a sequence of positive scalers $\{\alpha_n\}_{n=1}^{\infty}$ to update the score vector at time n. The counting variable $\nu_n(i,S,a)$ is equal to the number of times neighbor nodes S have received (and acknowledged) packets transmitted from node i and corresponding routing decision $a \in A(S)$ has been taken upto time n. Similarly, $N_n(i,S)$ is equal to the number of times set of nodes S have received (and acknowledged) packets transmitted from node i upto time n. Lastly, $\{\alpha_n\}_{n=1}^{\infty}$ is a fixed sequence of numbers available at all nodes.

Fig. 1 gives an overview of the components of the algorithm. Next we present further details.

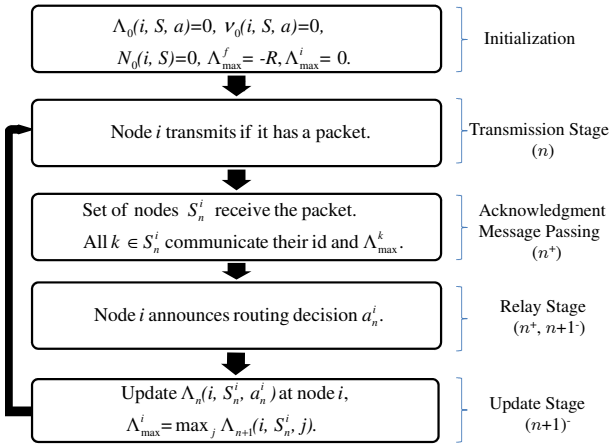


Fig. 1. Flow of the algorithm. The algorithm follows a four-stage procedure: transmission, acknowledgment, relay, and update.

B. Detailed description of d-AdaptOR

The operation of d-AdaptOR can be described in terms of initialization and four stages of transmission, reception and acknowledgment, relay, and adaptive computation as shown in Fig. 1. For simplicity of presentation we assume a sequential timing for each of the stages. We use n^+ to denote some (small) time after the start of n^{th} slot and $(n+1)^-$ to denote some (small) time before the end of n^{th} slot such that $n < n^+ < (n+1)^- < n+1$.

0) **Initialization**:
For all $i \in \Theta, S \in \mathfrak{S}^i, a \in A(S)$, initialize $\Lambda_0(i,S,a) = 0,\ \nu_0(i,S,a) = 0,\ N_0(i,S) = 0,\ \Lambda^f_{max} = -R, \Lambda^i_{max} = 0.$

1) **Transmission Stage**:
Transmission stage occurs at time n in which node i transmits if it has a packet.

2) **Reception and Acknowledgment Stage**:
Let S^i_n denote the (random) set of nodes that have received the packet transmitted by node i. In the reception and acknowledgment stage, successful reception of the packet transmitted by node i is acknowledged to it by all the nodes in S^i_n. We assume that the delay for the acknowledgment stage is small enough (not more than the duration of the time slot) such that node i infers S^i_n by time n^+.
For all nodes $k \in S^i_n$, the ACK packet of node k to node i includes the EBS message Λ^k_{max}.
Upon reception and acknowledgment, the counting random variable N_n is incremented as follows:

$$N_n(i,S) = \begin{cases} N_{n-1}(i,S)+1 & \text{if } S = S^i_n \\ N_{n-1}(i,S) & \text{if } S \neq S^i_n \end{cases}.$$

3) **Relay Stage**:
Node i selects a routing action $a^i_n \in A(S^i_n)$ according to the following (randomized) rule parameterized by $\epsilon_n(i,S) = \frac{1}{N_n(i,S)+1}$:

- with probability $(1-\epsilon_n(i,S^i_n))$,
$$a^i_n \in \arg\max_{j\in A(S^i_n)} \Lambda_n(i,S^i_n,j)$$
is selected,[2]
- with probability $\frac{\epsilon_n(i,S^i_n)}{|A(S^i_n)|}$,
$$a^i_n \in A(S^i_n)$$
is selected at random.

Node i transmits FO, a control packet which contains information about routing decision a^i_n at some time strictly between n^+ and $(n+1)^-$. If $a^i_n \neq f$, then node a^i_n prepares for forwarding in next time slot while nodes $j \in S^i_n, j \neq a^i_n$ expunge the packet. If termination action is chosen, i.e. $a^i_n = f$, all nodes in S^i_n expunge the packet.
Upon selection of routing action, the counting variable ν_n is updated.

$$\nu_n(i,S,a) = \begin{cases} \nu_{n-1}(i,S,a)+1 & \text{if } (S,a) = (S^i_n, a^i_n) \\ \nu_{n-1}(i,S,a) & \text{if } (S,a) \neq (S^i_n, a^i_n) \end{cases}.$$

4) **Adaptive Computation Stage**:
At time $(n+1)^-$, after being done with transmission and relaying, node i updates score vector $\Lambda_n(i,\cdot,\cdot)$ as follows:

- for $S = S^i_n,\ a = a^i_n$,
$$\Lambda_{n+1}(i,S,a) = \Lambda_n(i,S,a) + \alpha_{\nu_n(i,S,a)} \left(-\Lambda_n(i,S,a) + g(S,a) + \Lambda^a_{max} \right), \quad (2)$$
- otherwise,
$$\Lambda_{n+1}(i,S,a) = \Lambda_n(i,S,a). \quad (3)$$

Furthermore, node i updates its EBS message Λ^i_{max} for future acknowledgments as:

$$\Lambda^i_{max} = \max_{j\in A(S^i_n)} \Lambda_{n+1}(i,S^i_n,j).$$

C. Computational issues

The computational complexity and control overhead of d-AdaptOR is low.

[2]In case of ambiguity, node with the smallest index is chosen.

1) Complexity: To execute stochastic recursion (2), the number of computations required per packet is order of $O(\max_{i\in\Theta}|\mathcal{N}(i)|)$ at each time slot.

2) Control Overhead: The number of acknowledgments per packet is order of $O(\max_{i\in\Theta}|\mathcal{N}(i)|)$, independent of network size.

IV. Optimality of d-AdaptOR

We will now state the main result establishing the optimality of the proposed d-AdaptOR algorithm under an assumption of a time-invariant model of packet reception. More precisely, we have the following assumption.

Assumption 1. *The probability of successful reception of a packet transmitted by node i at set $S \subseteq \mathcal{N}(i)$ of nodes is $P(S|i)$, independent of time and all other concurrent transmissions.*

The probabilities $P(\cdot|\cdot)$ in Assumption 1 thus characterize a packet reception model which we refer to as *local broadcast model*. Note that for all $S \neq S'$, successful reception at S and S' are mutually exclusive and $\sum_{S\subseteq\Theta} P(S|i) = 1$. Furthermore, logically node i is always a recipient of its own transmission, i.e. $P(S|i) = 0$ if $i \notin S$.

The proposed local broadcast model is assumed to truly capture the coupling of the physical layer and the media access control (MAC) layer. In other words, the local broadcast model takes into account signal degradation due to path loss and multipath fading as well as captures the interference produced by other transmitting nodes. Note that, our model together with Assumption 1 imply an underlying MAC whose operation is controlled at a distinct layer and independently of the routing decisions. Furthermore, the implicit existence of a MAC scheme allows for a set of more advanced MAC schemes such as Zig-Zag [11]. Finally, the identically distributed assumption on successful transmissions imposes a time-homogeneity on the operation of the network and significantly restricts the topology changes of the network. In Sections V and VII, we address the severity and implications of the above consequences of Assumption 1. In particular, we will show that d-AdaptOR exhibits many of its desirable properties and performance improvements in practice despite relaxation of the analytical assumptions.

Let $\mathbb{P}$ be the sample space of the random probability measures for the local broadcast model. Specifically, $\mathbb{P} := \{p \in \mathbb{R}^{2^d} \times \mathbb{R}^d : \text{p is a non-square left stochastic matrix}\}$. Moreover, let $\mathcal{P}_P$ be the trivial σ-field generated by the local broadcast model $P \in \mathbb{P}$ (sample point in $\mathbb{P}$), i.e. $\mathcal{P}_P = \{P, \mathbb{P}\backslash P, \emptyset, \mathbb{P}\}$.[3] Let S_n^i be the set of nodes that have received the packet due to transmission from node i at time n, while a_n^i denotes the corresponding routing decision node i takes at time n.[4] A distributed routing policy is a collection $\phi = \{\phi^i\}_{i\in\Theta}$ of routing decisions taken at nodes $i \in \Theta$, where ϕ^i denotes a sequence of random actions $\phi^i = \{a_0^i, a_1^i, \ldots\}$ for node i. The policy ϕ is said to be *admissible* if for all nodes $i \in \Theta$, $S \in \mathfrak{S}^i$, $a \in A(S)$, the event $\{a_n^i = a\}$ belongs to the σ-field $\mathcal{H}_n^i$ generated by the observations at node i, i.e. $\bigcup_{j\in\mathcal{N}(i)}\{S_0^j, a_0^j, \ldots, S_{n-1}^j, a_{n-1}^j, S_n^j\}$. Let Φ denote the set of such admissible policies. These policies are implementable in a distributed manner under the following assumption.

Assumption 2. *The successful reception at set S due to transmission from node i is acknowledged perfectly to node i.*

With the above notations and assumptions, the following theorem establishes the optimality of d-AdaptOR, i.e. d-AdaptOR denoted by $\phi^* \in \Phi$, maximizes the expected average per packet reward obtained in (1) as $N \to \infty$.

Theorem 1. *Suppose $\sum_{n=0}^{\infty}\alpha_n = \infty$, $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$, and Assumptions 1 and 2 hold. Then for all $\phi \in \Phi$,*

$$\lim_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \geq \limsup_{N\to\infty} E^{\phi}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right]$$

where E^{ϕ^*} and E^{ϕ} are the expectations taken with respect to policies ϕ^* and ϕ respectively.[5]

Next we prove the optimality of d-AdaptOR in two steps. In the first step, we show that Λ_n converges in an almost sure sense. In the second step we use this convergence result to show that d-AdaptOR is optimal for Problem (**P**).

A. Convergence of Λ_n

Let $U : \prod_i \mathbb{R}^{v_i} \to \prod_i \mathbb{R}^{v_i}$ be an operator on vector Λ such that,

$$(U\Lambda)(i,S,a) = g(S,a) + \sum_{S'\in\mathfrak{S}^a} P(S'|a)\max_{j\in A(S')}\Lambda(a,S',j). \quad (4)$$

Let $\Lambda^* \in \prod_i \mathbb{R}^{v_i}$ denote the fixed point of operator U,[6] i.e.

$$\Lambda^*(i,S,a) = g(S,a) + \sum_{S'\in\mathfrak{S}^a} P(S'|a)\max_{j\in A(S')}\Lambda^*(a,S',j). \quad (5)$$

The following lemma establishes the convergence of recursion (2) to the fixed point of U, Λ^*.

Lemma 1. *Let*

(J1) $\Lambda_0(\cdot,\cdot,\cdot) = 0$, $\Lambda_{max}^f = -R$, $\Lambda_{max}^i = 0$ *for all* $i \in \Theta$,
(J2) $\sum_{n=0}^{\infty}\alpha_n = \infty$, $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$.

Then iterate Λ_n obtained by the stochastic recursion (2) converges to Λ^* almost surely.

The proof uses known results on the convergence of a certain recursive stochastic process as presented by Fact 3 in Appendix A.

[3] σ-field captures the knowledge of the realization of local broadcast model and assumes a well-defined prior on these models.

[4] $S_n^i = \emptyset, a_n^i = f$ if node i does not transmit at time n.

[5] This is a strong notion of optimality and implies that the proposed algorithm's expected average reward is greater than the best case performance (lim sup) of all policies [12, Page 344].

[6] Existence and uniqueness of Λ^* is provided in Appendix A.

B. Proof of optimality

Using the convergence of Λ_n we show that the expected average per packet reward under d-AdaptOR is equal to the optimal expected average per packet reward obtained for a genie-aided system where the local broadcast model is known perfectly.

In proving the optimality of d-AdaptOR, we take cue from known results associated with a closely related Auxiliary Problem **(AP)**. In the Auxiliary Problem **(AP)**, there exists a centralized controller with full knowledge of the local broadcast model $P(\cdot|\cdot)$ as well as the transmission outcomes across the network [1], [6]. For Auxiliary Problem **(AP)**, a routing policy is a collection $\pi = \{\pi^i\}_{i\in\Theta}$ of routing decisions taken for nodes $i \in \Theta$ at the centralized controller, where π^i denotes a sequence of random actions $\pi^i = \{a_0^i, a_1^i, \ldots\}$ for node i. The routing policy π is said to be admissible for Auxiliary Problem **(AP)** if the event $\{a_n^i = a\}$ belongs to the product σ-field $\mathcal{F}_n = \mathcal{P}_P \times \prod_i \mathcal{H}_n^i$ [13]. In this Auxiliary Problem **(AP)**, let Π denote the set of admissible policies for Auxiliary Problem **(AP)**. The reward associated with policy $\pi \in \Pi$ for routing a single packet m from the source to the destination is then given by

$$J^{\pi}(\{o\}) := \mathbf{E}^{\pi}\left[\left\{r_m - \sum_{n=0}^{\tau_e^m-1} c_{i_{n,m}}\right\}|\mathcal{F}_0\right], \qquad (6)$$

where $\mathcal{F}_0 = \mathcal{P}_P$. Now, in this setting, we are ready to formulate the following Auxiliary Problem **(AP)** as a classical shortest path Markov Decision Problem (MDP).

Auxiliary Problem (AP) Find an optimal policy π^* such that,

$$J^*(\{o\}) = J^{\pi^*}(\{o\}) = \sup_{\pi\in\Pi} J^{\pi}(\{o\}). \qquad (7)$$

Remark 1. The existence of an admissible policy $\pi^* \in \Pi$ achieving the supremum on the right hand side of (7) is a result of Theorem 7.1.9 in [12].

Auxiliary Problem **(AP)** has been extensively studied in [1], [6], [14] and the following theorem has been established in [6].

Fact 1. [6, Theorem 2.1] There exists a function $\bar{\pi}^* : \{\mathfrak{S}^i\}_{i\in\Theta} \to \Theta \cup \{f\}$ such that $\pi^* = \{a_0^i, a_1^i, \ldots\}_{i\in\Theta}$ is an optimal policy for Auxiliary Problem **(AP)**, where $a_n^i = \bar{\pi}^*(S_n^i)$.[7]

Furthermore, $\bar{\pi}^*$ is such that

$$\bar{\pi}^*(S) \in \arg\max_{j\in A(S)} V^*(j), \qquad (8)$$

where (value) function $V^* : \Theta \cup \{f\} \to \mathbb{R}^+$ is the unique solution to the following fixed point equation:

$$V^*(d) = R \qquad (9)$$

$$V^*(i) = \max(\{-c_i + \sum_{S'} P(S'|i)(\max_{j\in S'} V^*(j))\}, 0) \qquad (10)$$

$$V^*(f) = 0. \qquad (11)$$

[7]In other words there exists a stationary, deterministic, and Markov optimal policy for Auxiliary Problem **(AP)**.

Moreover, $V^*(j)$ is the maximum expected reward for routing a packet from node j to destination d, i.e.

$$V^*(j) = \sup_{\pi\in\Pi} J^{\pi}(\{j\}).$$

Lastly,

Fact 2. [14, Proposition 4.3.3] function $V^* : \Theta \cup \{f\} \to \mathbb{R}^+$ is unique.

Lemma 2 below states the relationship between the solution of Problem **(P)** and that of the Auxiliary Problem **(AP)**. More specifically, Lemma 2 shows that $V^*(o)$ is an upper bound for the solution to Problem **(P)**.

Lemma 2. *Consider any admissible policy $\phi \in \Phi$ for Problem* **(P)**. *Then for all $N = 1, 2, \ldots,$*

$$E^{\phi}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \le V^*(o).$$

Proof: The proof is given in Appendix B. Intuitively the result holds because the set of admissible policies in **(P)** is a subset of admissible policies in **(AP)**, i.e. $\Phi \subset \Pi$. ■

Lemma 3 gives the achievability proof for Problem **(P)** by showing that the expected average per packet reward of d-AdaptOR is no less than $V^*(o)$.

Lemma 3. *For any $\delta > 0$,*

$$\liminf_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \ge V^*(o) - \delta.$$

Proof: The proof is given in Appendix C. ■

Lemmas 2 and 3 imply that

$$\lim_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right]$$

exists and is equal to $V^*(o)$ establishing the proof of Theorem 1.

V. Protocol Design and implementation issues

In this section we describe an 802.11 compatible implementation for d-AdaptOR.

A. 802.11 compatible implementation

Implementation of d-AdaptOR, analogous to any opportunistic routing scheme involves the selection of a relay node from a candidate set of nodes that have received and acknowledged a packet successfully. One of the major challenges in devising d-AdaptOR algorithm is the design of 802.11 compatible acknowledgment mechanism at the MAC layer. Below we propose a practical and simple to implement acknowledgment architecture.

For each neighbor node $j \in \mathcal{N}(i)$, the transmitter node i reserves a virtual time slot of duration $T_{ACK} + T_{SIFS}$, where T_{ACK} is the duration of the acknowledgment packet and T_{SIFS} is the duration of Short InterFrame Space (SIFS) [15]. The transmitter i then piggy-backs a priority ordering of

nodes $\mathcal{N}(i)$ with each data packet transmitted. The priority ordering determines the virtual time slot in which a candidate node transmits an acknowledgment. Nodes in the set S^i that have successfully received the packet then transmit acknowledgment packets sequentially in the reserved virtual time slots in the order determined by the transmitter node. For example, in the linear network shown in Fig. 2, if node o piggy-backs the order {1,2}, then node 1 transmits an ACK first and later node 2 transmits an ACK. If node 1 does not receive the packet successfully from node o, node 1 does not transmit an ACK and a duration of $T_{ACK} + T_{SIFS}$ corresponding to node 1 is not utilized.

For receiving ACKs, each transmitting node i waits for a duration of $T_{wait} = |\mathcal{N}(i)|(T_{ACK} + T_{SIFS})$. After each node in the set S^i has acknowledged or T_{wait} timer has expired, node i transmits a FOrwarding control packet (FO). If timer T_{wait} has expired and no ACK has been received, then node i either drops the packet or retransmits. If priority of node $j \in S^i$ is l, $1 \leq l \leq |\mathcal{N}(i)|$, then it waits for a duration of $T_{waitFO} = (|\mathcal{N}(i)| - l + 1)(T_{ACK} + T_{SIFS})$ to receive a FO. If T_{waitFO} expires and no FO packet has been received, then the corresponding candidate nodes drop the received data packet. Fig 3 shows a typical sequence of control packets for topology in Fig 2.

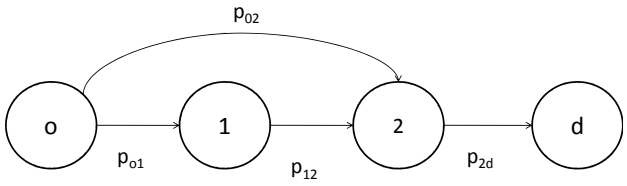


Fig. 2. With probability p_{ij}, a packet transmitted by node i is successfully received by node j

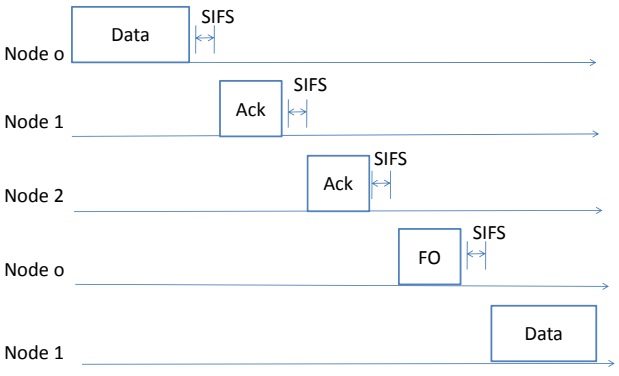


Fig. 3. Typical packet transmission sequence for d-AdaptOR

In addition to the acknowledgment scheme, d-AdaptOR requires modifications to the 802.11 MAC frame format. Fig. 4 shows the modified MAC frame formats required by d-AdaptOR. The reserved bits in the type/subtype fields of the frame control field of the 802.11 MAC specification are used to indicate whether the rest of the frame is a d-AdaptOR data frame, a d-AdaptOR ACK, or a FO. This enables the d-AdaptOR to communicate and be fully compatible with other 802.11 devices.

The data frame contains the candidate set in priority order, the payload, and the 802.11 Frame Check Sequence. The acknowledgment frame includes the data frame senders address and the feedback EBS Λ_{max}. The FO packet is exactly the same as a standard 802.11 short control frame, but uses different subtype value.

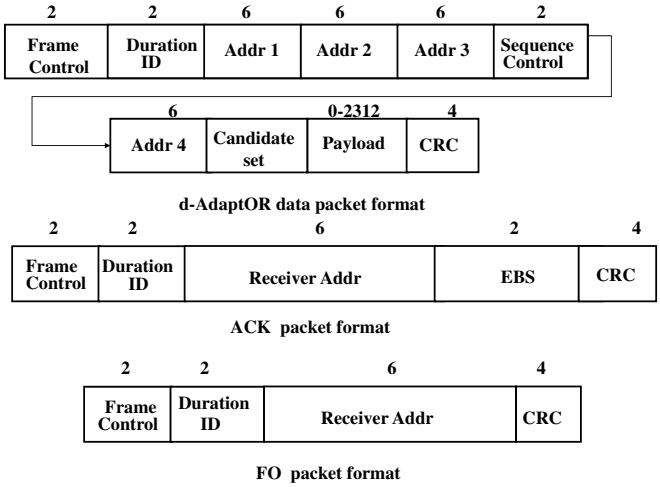


Fig. 4. Frame structure of the data packet, acknowledgment packet, and FO packet

B. d-AdaptOR in non-idealized setting

1) Loss of ACK and FO packets: Interference or low SNR can cause loss of ACK and FO packets. Loss of an ACK packet results in an incorrect estimation of nodes that have received the packet and thus affects the performance of the algorithm. Loss of FO packet negatively impacts the throughput performance of the network. In particular, loss of FO packet can result in the drop of data packet at all the potential relays, reducing the throughput performance.

2) Increased Overhead: d-AdaptOR adds a modest additional overhead to standard 802.11 due to the added acknowledgment/handshake structure. Assuming a 802.11b physical layer operating at 11 Mbps with a SIFS time of 10 μs, preamble duration of 20 μs, Physical Layer Convergence Protocol (PLCP) header duration of 4 μs and 512 byte frame payloads, the overhead of an d-AdaptOR data frame with three candidates is compared with unicast 802.11 in Table I. It is clear that the overhead increases linearly with the number of neighbors.

TABLE I
OVERHEAD COMPARISONS

	Data Frame	ACK	Total
802.11 unicast	397 μs	46 μs	443 μs
d-AdaptOR	400 μs	125 μs	525 μs

Note that the overhead cost can be reduced by restricting the number of nodes in the candidate list of MAC header to a given number, MAX-NEIGHBOUR. The unique ordering for the nodes in the candidate set is determined by prioritizing the nodes with respect to $\Lambda_n(i,\{i,j\},j), j \in \mathcal{N}(i)$ and then choosing the MAX-NEIGHBOUR highest priority nodes.[8] Needless to say that such limitation will sacrifice the optimality of d-AdaptOR for a lower overhead.

3) Choice of parameters: To ensure an acceptable throughput, the value of reward, R, must be chosen sufficiently high. However, beyond a given threshold (depending on the network topology), the value of R does not affect asymptotic performance of the algorithm. The convergence rate of stochastic recursion (2) strongly depends on choice of sequence $\{\alpha_n\}$. It converges slowly with slowly decreasing sequence $\{\alpha_n\}$ and results in less variance in the estimates of Λ_n, while fast decreasing sequence of $\{\alpha_n\}$ causes large variance in the estimates of Λ_n.

VI. Simulation Study

In this section, we provide simulation results in which the performance of d-AdaptOR is compared against suitably chosen candidates: Stochastic Routing (SR) [1] (SR is the distributed implementation of policy π^* discussed in Section IV-B), EXOR [4] and a conventional routing algorithm Ad hoc On-Demand Distance Vector Routing (AODV) [15]. Both SR and EXOR are distributed mechanisms in which the probabilistic structure of the network is used to implement opportunistic routing algorithms. As a result, their performance will be highly dependent on the precision of empirical probability associated with link, p_{ij}. In fact, authors in [8] have identified network topologies and examples in which small errors in empirical probabilities incur significant loss of performance. To provide a fair comparison, hence, we have considered modified versions of SR and EXOR in which the algorithms adapt p_{ij} to the history of packet reception outcomes, while rely on the updates to make routing decisions (separated scheme of estimation and routing).

Our simulations are performed in QualNet. Simulations consist of a grid topology of 16 nodes as shown in Fig. 5(a) each equipped with 802.11b radios transmitting at 11 Mbps. The wireless medium is modeled as to include Rician fading and Log-normal shadowing with mean 4dB and the path loss follows the two-ray model in [16] with path exponent of 3. Note that the choice of indoor environment is motivated by the findings in [17] where opportunistic routing is found to provide better diversity of transmission outcomes.

Packets are generated according to a CBR source with rate 10 packets/sec. They are assumed to be of length 512 bytes equipped with simple CRC error detection. The acknowledgment packets are short packets of length 24 bytes transmitted at rate of 11 Mbps, while FO packets are transmitted at reliable lower rate of 1Mbps. The lower transmission rate for FO packets is used as it increases the reliability of the packets to avoid issues discussed in Section V-B. Cost of transmission is assumed to be one unit, while reward for successfully delivering a packet to the destination is assumed to be 20.

Fig. 5(b) plots the expected average per packet reward obtained by the candidate routing algorithms versus network operation time. The optimal algorithm with complete knowledge of link probabilities (presented in Section IV-B) is also plotted for comparison. We first note that as expected, ADOV performs poorly compared to the opportunistic schemes as it is strictly suboptimal. In particular, Fig. 5(b) shows that the d-AdaptOR algorithm outperforms opportunistic schemes EXOR and SR by at least 5% given sufficient number of packet deliveries. Fig. 5(b) shows that SR performs poorly relative to d-AdaptOR algorithm since it fails to explore possible choices of routes and often results in strictly suboptimal routing policy.

This figure also shows that the randomized routing decisions employed by d-AdaptOR work as a double-edge sword. This is the mechanism through which network opportunities are exhaustively explored until the globally optimal decisions are constructed. At the same time, these randomized decisions lead to a short term performance loss. One should note that due to the exploratory nature of the d-AdaptOR algorithm, during initial startup time EXOR and SR perform better than d-AdaptOR. This in fact is reminiscent of the well-known exploration/exploitation trade-off in stochastic control and learning literature.

To clearly manifest the differences in the performance on the rate of convergence, one can define a finite horizon quantity, *regret*, as [18]

$$Re(N) = \mathbf{E}\left[\sum_{m=1}^{M_N} V^*(o) - \left(r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}}\right)\right]. \quad (12)$$

Regret demonstrates the performance of a routing policy over a finite horizon. Hence, it provides an appropriate measure of comparison in scenarios with finitely many packets. Fig. 5(c) again shows that d-AdaptOR outperforms EXOR and SR after sufficient time.

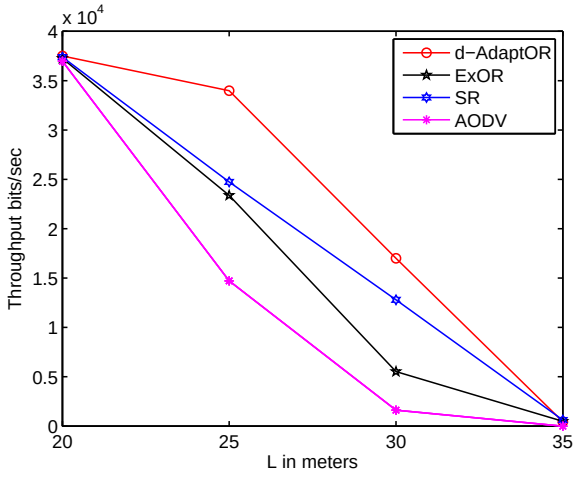
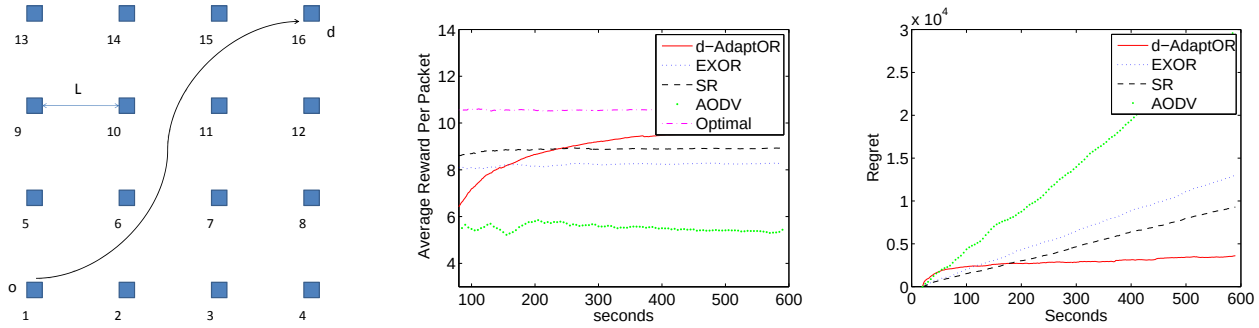


Fig. 6. Throughput comparisons for d-AdaptOR, SR, EXOR, AODV

Throughput is a benchmark criterion to measure net-

[8] In case of ambiguity, node with the smallest index is chosen.



(a) Node topology (b) Expected Average Reward (c) Regret

Fig. 5. d-AdaptOR vs distributed SR, EXOR, and AODV. Parameters: L=25 meters, $\alpha_n = \frac{1}{n\log(n)}$, $R = 20$, $c_i = 1$ for all i.

work performance.[9] Fig. 6 compares the average throughput obtained by the three opportunistic routing algorithms (d-AdaptOR, EXOR, and SR) and AODV as the distance L in Fig. 2 is varied. Note that for low and high values of L, the diversity is low while medium values of L provide high level of diversity (for low values of L network becomes single hop while for large values of L network gets disconnected). Opportunistic routing schemes perform better as compared to the conventional routing when the network provides high diversity.

Finally, we investigate the performance result for a random topology in Fig. 7(a), wherein 16 nodes are uniformly distributed over an area of 90m $\times$ 90m and all other parameters are kept the same as those in the grid topology of Fig. 5(a). Fig. 7(b), 7(c) plots the average reward and regret respectively for the candidate routing algorithms. The results are inline with conclusions for the grid topology in Fig. 5(a). It should be noted that, optimality of d-AdaptOR holds for all topologies, however it may not outperform other algorithms (SR, EXOR) in certain topologies.

VII. Conclusions

In this paper, we proposed d-AdaptOR, an adaptive routing scheme which maximizes the expected average per packet reward from a source to a destination in the absence of any knowledge regarding network topology and link qualities. d-AdaptOR allows for a practical distributed implementation with provably optimal performance under idealized assumptions on stationarity of network and reliability of acknowledgment scheme.

The performance of d-AdaptOR is also investigated in practical settings via simulations. Simulation results show that d-AdaptOR outperforms the existing opportunistic protocols in which statistical link qualities are empirically built and the routing decisions are greedily adapted to the empirical link models.

The long term average reward criterion investigated in this paper is somewhat limited in discriminating among various adaptive schemes with optimal average reward per packet. This is mostly due to the inherent dependency of the long term average reward on the tail events. To capture the performance of various adaptive schemes e.g. convergence rate, it is desirable to study the regret as defined in (12). An important area of future work comprises of developing fast converging algorithms which optimize the regret as a performance measure of interest.

The design of routing protocols need consideration of congestion control along with the throughput performance [19], [20]. Our work, however does not consider the issue of congestion control. Incorporating congestion control in opportunistic routing algorithms to minimize expected delay in an oblivious network is an area of future research.

Last but not least, the broadcast model used in this paper assumes a decoupled operation at the MAC and network layer. While this assumption seems reasonable for many popular MAC schemes based on random access philosophy, it ignores the potentially rich interplays between scheduling and routing which arise in scheduling TDMA-based schemes.

[9]Throughput is directly related to the quantity defined in (1) for stable arrival rates and transmission cost as hop count.

Appendix

A. Proof of Lemma 1

Lemma 1. Let

(J1) $\Lambda_0(\cdot,\cdot,\cdot) = 0$, $\Lambda^f_{max} = -R$, $\Lambda^i_{max} = 0$ for all $i \in \Theta$,
(J2) $\sum_{n=0}^{\infty} \alpha_n = \infty$, $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$.

Then iterate Λ_n obtained by the stochastic recursion (2)

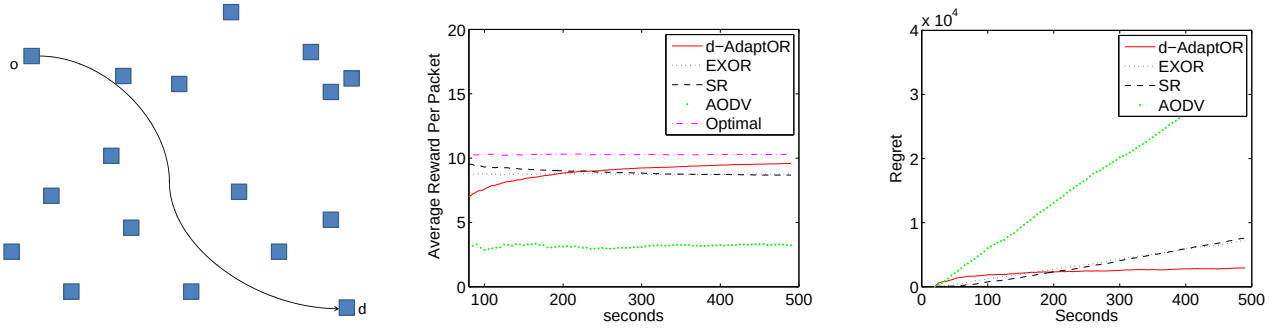
$$\Lambda_{n+1}(i,S,a) = \Lambda_n(i,S,a) + \alpha_{\nu_n(i,S,a)} \left(-\Lambda_n(i,S,a) + g(S,a) + \Lambda^a_{max} \right),$$

converges to Λ^* almost surely.

To prove Lemma 1, we note that the adaptive computation given by (2) utilizes a stochastic approximation algorithm to solve the MDP associated with Problem **(AP)**. To study the convergence properties of this stochastic approximation, we appeal to known results in the intersection of learning and stochastic approximation given below.

In particular, consider a set of stochastic sequences on $\mathbb{R}^D$, denoted by $\{x_n, \bar{\alpha}_n, \mathcal{M}_{n+1}\}$, and the corresponding filtration $\mathcal{G}_n$, i.e. the increasing σ-field generated by $\{x_n, \bar{\alpha}_n, \mathcal{M}_{n+1}\}$, satisfying the following recursive equation

$$x_{n+1} = x_n + \bar{\alpha}_n[U(x'_n) - x_n + \mathcal{M}_{n+1}],$$



(a) Node topology : Nodes placed uniformly over an area of 90m × 90m

(b) Expected Average Reward

(c) Regret

Fig. 7. d-AdaptOR vs distributed SR, EXOR, and AODV. $\alpha_n = \frac{1}{n\log(n)}$, $R = 20$, $c_i = 1$ for all i.

where U is a mapping from $\mathbb{R}^D$ into $\mathbb{R}^D$ and $x'_n = \left(x_{n_1}(1), x_{n_2}(2), ..., x_{n_D}(D)\right), 0 \le n_j \le n,\ j \in \{1, 2, ..., D\}$, is a vector of possibly delayed components of x_n. If no information is outdated, then $n_j = n$ for all j and $x'_n = x_n$. The following important result on the convergence of x_n is provided in [21].

Fact 3 (Theorem 1, Theorem 2 [21]). Assume $\{x_n, \bar{\alpha}_n, \mathcal{M}_{n+1}\}$ and U satisfy the following conditions:

(G1) For all $n \ge 0$ and $1 \le l \le D$, $0 \le \bar{\alpha}_n(l) \le 1$ a.s.; for $1 \le l \le D$, $\sum_{n=0}^{\infty} \bar{\alpha}_n(l) = \infty$ a.s.; for $1 \le l \le D$, $\sum_{n=0}^{\infty} \bar{\alpha}_n^2(l) < \infty$ a.s.

(G2) $\mathcal{M}_n$ is a martingale difference with finite second moment, i.e. $\mathbf{E}\{\mathcal{M}_{n+1}|\mathcal{G}_n\} = 0$, and there exist constants A and B such that $\mathbf{E}\{\mathcal{M}_{n+1}^2|\mathcal{G}_n\} \le A + B(\max_{n' \le n} ||x_{n'}||)^2$.

(G3) There exists a positive vector v, scalers $\beta \in [0, 1)$ and $C \in \mathbb{R}^+$, such that

$$||U(x)||_v \le \beta ||x||_v + C.$$

(G4) Mapping $U : \mathbb{R}^D \to \mathbb{R}^D$ satisfies the following properties:

1) U is component-wise monotonically increasing;
2) U is continuous;
3) U has a unique fixed point $x^* \in \mathbb{R}^D$;
4) $U(x) - r\mathbf{1} \le U(x - r\mathbf{1}) \le U(x + r\mathbf{1}) \le U(x) + r\mathbf{1}$, for any $r \in \mathbb{R}^+$.

(G5) For any j, $n_j \to \infty$ as $n \to \infty$.

Then the sequence of random vectors x_n converges to the fixed point x^* almost surely.

Let $\mathcal{G}_n$ be the increasing σ-field generated by random vectors $(\Lambda_n, S_n^i, a_n^i, \nu_n)$. Let $x_n = \Lambda_n$ be the random vector of dimension $D = \sum_{i\in\Theta} \sum_{S\in\mathfrak{S}^i} |A(S)|$, generated via recursive equation (2). Furthermore,

$$(U\Lambda_n)(i, S, a) = g(S, a) + \sum_{S' \in \mathfrak{S}^a} P(S'|a) \max_j \Lambda_n(a, S', j),$$

$$\bar{\alpha}_n(i, S, a) = \alpha_{\nu_n(i,S,a)} I(S_n^i = S, a_n^i = a).$$

Let $\{\mathcal{M}_{n+1}\}$ be a random vector whose $(i, S, a)^{th}$ element is constructed as follows:

$$\begin{aligned}\mathcal{M}_{n+1}(i, S, a) &= \max_j \Lambda_{n_a}(a, S_{n_a}^a, j) \\ &\quad - \sum_{S' \in \mathfrak{S}^a} P(S'|a) \max_j \Lambda_{n_a}(a, S', j),\end{aligned}$$

where $0 \le n_a \le n$, and $S_{n_a}^a$ is the most recent state visited by node a.

Now we can rewrite (2) and (3) as in the form investigated in Fact 3, i.e.

$$\begin{aligned}\Lambda_{n+1}(i, S, a) &= \Lambda_n(i, S, a) + \bar{\alpha}_n(i, S, a)\Big((U\Lambda_{n_a})(i, S, a) \\ &\quad - \Lambda_n(i, S, a) + \mathcal{M}_{n+1}(i, S, a)\Big).\end{aligned}$$

The remaining steps of the proof reduces to verifying statements (G1)-(G5). This is verified in Lemma 4 below.

Lemma 4. *$(\Lambda_n, \bar{\alpha}_n, \mathcal{M}_{n+1})$ satisfy conditions (G1)-(G5).*

Proof:

- (G1): It is shown in Lemma 6 that algorithm d-AdaptOR guarantees that every state-action is attempted infinitely often (i.o.). Hence,

$$\begin{aligned}\sum_{n=0}^{\infty} \bar{\alpha}_n(i, S, a) &= \sum_{n=0}^{\infty} \alpha_{\nu_n(i,S,a)} I(S_n^i = S, a_n^i = a) \\ &\ge I((i, S, a) \text{ visited i.o. }) (\sum_{n=0}^{\infty} \alpha_n) = \infty.\end{aligned}$$

However,

$$\begin{aligned}\sum_{n=0}^{\infty} \bar{\alpha}_n^2(i, S, a) &\le \sum_{i,S,a} \sum_{n=0}^{\infty} \alpha_{\nu_n(i,S,a)}^2 I(S_n^i = S, a_n^i = a) \\ &\le \sum_{i\in\Theta} |\mathfrak{S}^i||d + 1| \sum_{n=0}^{\infty} \alpha_n^2 < \infty.\end{aligned}$$

- (G2):

$$\begin{aligned}\mathbf{E}[\mathcal{M}_{n+1}|\mathcal{G}_n, n_a] &= \mathbf{E}_{S^a}[\max_j \Lambda_{n_a}(a, S^a, j)] \\ &\quad - \sum_{S'} P(S'|a) \max_j \Lambda_{n_a}(a, S', j) \\ &= 0.\end{aligned}$$

$$\mathbf{E}[\mathcal{M}_{n+1}|\mathcal{G}_n] = \mathbf{E}_{n_a}\left[\mathbf{E}[\mathcal{M}_{n+1}|\mathcal{G}_n, n_a]\right] = 0.$$

$$\begin{aligned}\mathbf{E}[\mathcal{M}_{n+1}^2|\mathcal{G}_n, n_a] &\leq \mathbf{E}_{S^a}[(\max_j \Lambda_{n_a}(a, S^a, j))^2] \\ &\leq \max_{S^a}\max_j \left(\Lambda_{n_a}(a, S^a, j)\right)^2 \\ &\leq ||\Lambda_{n_a}||^2.\end{aligned}$$

$$\begin{aligned}\mathbf{E}[\mathcal{M}_{n+1}^2|\mathcal{G}_n] &= \mathbf{E}_{n_a}\left[\mathbf{E}[\mathcal{M}_{n+1}^2|\mathcal{G}_n, n_a]\right] \\ &\leq \mathbf{E}_{n_a}[||\Lambda_{n_a}||^2] \\ &\leq \max_{n'\leq n}||\Lambda_{n'}||^2.\end{aligned}$$

Therefore Assumption (G2) of Fact 3 is satisfied.

- (G3): Let $Z_d = \left\{S : d \in S, S \in \{\mathfrak{S}^i\}_{i\in\Theta}\right\}$ denote the set of states which contain the destination node d. Moreover, let $Z_d^i = \{S : d \in S, i \in \Theta, S \in \mathfrak{S}^i\}$. Let $\tau_{Z_d}^\pi$ be the hitting time associated with set Z_d and policy $\pi \in \Pi$, i.e. $\tau_{Z_d}^\pi = \min\left\{n > 0 : \exists S \in Z_d, S \in \{S_n^i\}_{i\in\Theta}\right\}$. Policy π is said to be proper if $Prob(\tau_{Z_d}^\pi < \infty|\mathcal{F}_0) = 1$. Let us now fix a proper deterministic stationary policy $\pi \in \Pi$. Existence of such a policy is guaranteed from the connectivity between o and d. Let F be the termination state which is reached after taking the termination action f. Let us define a policy dependent operator $\mathcal{L}^\pi$,

$$(\mathcal{L}^\pi\Lambda)(i, S, a) = g(S, a) + \sum_{S' \notin Z_d^a \cup F} P(S'|a)\Lambda(a, S', \pi(S')). \tag{13}$$

We then consider a Markov chain with states (i, S, a) and with the following dynamics: from any state (i, S, a), we move to state $(a, S', \pi(S'))$, with probability $P(S'|a)$. Thus, subsequent to the first transition, we are always at a state of the form $(i, S, \pi(S))$ and the first two components of the state evolve according to policy π. As π is assumed proper, it follows that the system with states (i, S, a) also evolves according to a proper policy. We construct a matrix Q with each entry corresponding to the transition from state (i, S) to $(\pi(S), S')$ with value equal to $P(S'|\pi(S))$ for all $S \notin Z_d^i \cup F, S' \notin Z_d^{\pi(S)} \cup F$ for all i.
Since policy π is proper, the maximum eigenvalue of matrix Q is strictly less than 1. As Q is a non-negative matrix, Perron Frobenius theorem guarantees the existence of a positive vector w with components $w_{(i,S,a)}$ and some $\beta \in [0, 1)$ such that

$$\sum_{S' \notin Z_d^a \cup F} P(S'|a)w_{(\pi(S),S',\pi(S'))} \leq \beta w_{(i,S,a)}. \tag{14}$$

From (14), we have a positive vector v such that $||(\mathcal{L}^\pi\Lambda) - \Lambda^\pi||_v \leq \beta\,||\Lambda - \Lambda^\pi||_v$, where Λ^π is the fixed point of equation $\Lambda = \mathcal{L}^\pi\Lambda$.
From the definition of U (4) and $\mathcal{L}^\pi$ (13) we have $|(U\Lambda)(\cdot,\cdot,\cdot)| \leq |(\mathcal{L}^\pi\Lambda)(\cdot,\cdot,\cdot)|$. Using this and the triangle inequality, we obtain

$$\begin{aligned}||U\Lambda||_v &\leq ||\mathcal{L}^\pi\Lambda||_v \\ &\leq ||\mathcal{L}^\pi\Lambda - \mathcal{L}^\pi\Lambda^\pi||_v + ||\mathcal{L}^\pi\Lambda^\pi||_v \\ &\leq \beta\,||\Lambda - \Lambda^\pi||_v + ||\Lambda^\pi||_v \\ &\leq \beta\,||\Lambda||_v + 2\,||\Lambda^\pi||_v,\end{aligned}$$

establishing the validity of (G3).

- (G4): Assumption (G4) is satisfied by operator U using following fact:

  Fact 4. [Proposition 4.3.1 [14]] U is monotonically increasing, continuous, and satisfies $U(\Lambda) - r\mathbf{1} \leq U(\Lambda - r\mathbf{1}) \leq U(\Lambda + r\mathbf{1}) \leq U(\Lambda) + r\mathbf{1}$, $r > 0$.

  Λ^* is a fixed point of U. From (5) and (9)-(11) we obtain

$$\max_{j\in A(S)} V^*(j) = \max_{j\in A(S)} \Lambda^*(i, S, j) + R. \tag{15}$$

  Furthermore, using (5) and (15), for all $i \in \Theta$

$$\Lambda^*(i, S, a) = g(S, a) + \sum_{S'} P(S'|a)\max_{j\in A(S')} V^*(j) - R. \tag{16}$$

  The existence of fixed point Λ^* follows from (16), while uniqueness of Λ^* follows from uniqueness of V^* (Fact 2).

- (G5): Suppose $n_a \nrightarrow \infty$ as $n \to \infty$. Therefor, there exists N such that $n_a < N$ for all n. This means that the number of times that node a has transmitted a packet is bounded by N. But this contradicts Lemma 6 which says that each state-action pair (S, a) is visited i.o. Therefore $n_a \to \infty$ as $n \to \infty$ for all a, and condition (G5) holds.

Thus Assumptions (G1)-(G5) are satisfied. Hence, from Fact 3 our iterate (2) converges almost surely to Λ^*, the unique fixed point of U. ■

Lemma 5. *If policy ϕ^* is followed, then action $a \in A(S)$ is selected i.o. if state $S \in \mathfrak{S}$ is visited i.o.*

Proof:

Define random variable $K_n = I(S_n^i = S$ for any $i \in \Theta|\phi^*)$. Let $\mathcal{K}_n$ be the σ-field generated by $(K_1, K_2, \ldots, K_n)$. Let $A_n = \{\omega : a_n^i = a, S_n^i = S$ for any $i \in \Theta|\phi^*\}$.

From the construction of the algorithm it is clear that A_n is $\mathcal{K}_n$ measurable. Now it is clear that under policy ϕ^*, A_{n+1} is independent of K^{n-1} given K_n and $N_n(i, S), i \in \Theta$. Define,

$$P(A_{n+1}|K_n, N_n(i, S) \text{ for all } i \in \Theta) \geq \begin{cases} 0 & \text{if } K_n = 0 \\ \frac{\min_{i\in\Theta}\epsilon_n(i,S)}{|A(S)|} & \text{if } K_n = 1 \end{cases} \tag{17}$$

$$\begin{aligned}
&\sum_{n=0}^{\infty} Prob(A_{n+1}|\mathcal{K}_n) \\
&\geq \sum_{n=0}^{\infty} Prob(A_{n+1}|K_n, N_n(i,S) \text{ for all } i \in \Theta) \\
&\geq I(\text{ S is visited i.o.}) \sum_{n=0}^{\infty} \min_{i\in\Theta} \frac{\epsilon_n(i,S)}{|A(S)|} \\
&\geq \frac{I(\text{ S is visited i.o.})}{|A(S)|} \sum_{n=0}^{\infty} \frac{1}{\sum_{i\in\Theta} N_n(i,S)+1} \\
&\geq \frac{I(\text{ S visited i.o. })}{|A(S)|} \sum_{n=0}^{\infty} \frac{1}{n(d+1)+1} = \infty. \quad (18)
\end{aligned}$$

Next step of the proof is based on the following fact.

Fact 5 (Corollary 5.29, [22], (Extended Borel-Cantelli Lemma)). Let $\mathcal{K}_k$ be an increasing sequence of σ-fields and let A_k be $\mathcal{K}_k$-measurable. If $\sum_{k=1}^{\infty} Prob(A_k|\mathcal{K}_{k-1}) = \infty$ then $P(A_k \ i.o.) = 1$.

Thus, from Fact 5, $a \in A(S)$ is visited i.o, if S is visited i.o. ∎

Lemma 6. *If policy ϕ^* is followed, then each state-action (S,a) is visited infinitely often.*

Proof: We say states $S, S' \in \mathfrak{S}$ communicate if there exists a sequence of actions $\{a_1, \dots a_k, k < \infty\}$ such that probability of reaching state S' from state S following the sequence of actions $\{a_1, \dots a_k\}$ is greater than zero. Using Lemma 5, if state $S \in \mathfrak{S}$ is visited i.o., then every action $a \in A(S)$ is chosen i.o. as set $A(S)$ is finite. Hence, states S' such that $P(S'|a) > 0, S' \in \mathfrak{S}$, are visited i.o. if S is visited i.o. By Lemma 5 every action $a' \in A(S')$ is also visited i.o. Following similar argument and repeated application of Lemma 5, every state $S'' \in \mathfrak{S}$, which communicates with state S and actions $a \in A(S'')$ are visited i.o.

Under the assumption of the packet generation process in Section II, a packet is generated i.o. at the source node o. Thus state $\{o\}$ is reached i.o. The construction of set $\mathfrak{S}$ is such that every state $S \in \mathfrak{S}$ communicates with state $\{o\}$. Thus each (S,a) is visited i.o since $|\mathfrak{S}|$ is finite. ∎

B. Proof of Lemma 2

Lemma 2. Consider any admissible policy $\phi \in \Phi$ for Problem $(\mathbf{P})$. Then

$$E^{\phi}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \leq V^*(o).$$

Proof: Since π^* is the optimal policy for one packet, for each packet m and for any feasible policy $\phi \in \Phi$,

$$\begin{aligned}
V^*(o) &= \mathbf{E}^{\pi^*}\left[r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}|\mathcal{F}_0\right] \\
&\geq \mathbf{E}^{\phi}\left[r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right].
\end{aligned}$$

$$\begin{aligned}
&\mathbf{E}^{\phi}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \\
&\leq \mathbf{E}^{\phi}\left(\frac{1}{M_N}\sum_{m=1}^{M_N} V^*(o)\right) \\
&= V^*(o).
\end{aligned}$$

∎

C. Proof of Lemma 3

Lemma 3. For any $\delta > 0$,

$$\liminf_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \geq V^*(o) - \delta.$$

Proof: From (5), (9)-(11), and (15) we obtain the following equality for all $i \in \Theta,\ S \in \mathfrak{S}^i$,

$$\arg\max_{j\in A(S)} V^*(j) = \arg\max_{j\in A(S)} \Lambda^*(i,S,j)\ . \quad (19)$$

Let

$$b = \min_{i\in\Theta}\min_{S\in\mathfrak{S}^i} \min_{\substack{j,k\in A(S) \\ \Lambda^*(i,S,j)\neq\Lambda^*(i,S,k)}} |\Lambda^*(i,S,j) - \Lambda^*(i,S,k)|.$$

Lemma 1 implies that, in an almost sure sense, there exists packet index $m_1 < \infty$ such that for all $n > \tau_s^{m_1}$, $i \in \Theta$, $S \in \mathfrak{S}^i$, $a \in A(S)$,

$$|\Lambda_n(i,S,a) - \Lambda^*(i,S,a)| \leq b/2.$$

In other words, from time $\tau_s^{m_1}$ onwards, given any node $i \in \Theta$ and set $S \in \mathfrak{S}^i$, the probability that d-AdaptOR chooses an action $a \in A(S)$ such that $\Lambda^*(i,S,a) \neq \max_{j\in A(S)} \Lambda^*(i,S,j)$ is upper bounded by $\epsilon_n(i,S)$. Furthermore, since $N_n(i,S) \to \infty$ (Lemma 6), for a given $\gamma > 0$, with probability 1, there exists a packet index $m_2 < \infty$ such that for all $n > \tau_s^{m_2}$, $\max_{i,S} \epsilon_n(i,S) < \gamma$.

Let $m_0 = \max\{m_1, m_2\}$. For all packets with index $m \leq m_0$ the overall expected reward is upper-bounded by $m_0 R < \infty$ and lower-bounded by $-\frac{m_0}{\lambda} d \max_i c_i > -\infty$, hence, their presence does not impact the expected average per packet reward. Consequently, we only need to consider the routing decisions of policy ϕ^* for packets $m > m_0$.

Consider the m^{th} packet generated at the source. Let B_k^m be an event for which there exist k instances when d-AdaptOR routes packet m differently from the possible set of optimal actions. Mathematically speaking, event B_k^m occurs iff there exists instances $\tau_s^m \leq n_1^m \leq n_2^m \dots n_k^m \leq \tau_e^m$ such that for all $l = 1, 2, \dots, k$

$$\Lambda^*(i_{n_l^m,m}, S_{n_l^m}, a_{n_l^m}) \neq \max_{j\in A(S_{n_l^m})} \Lambda^*(i_{n_l^m,m}, S_{n_l^m}, j),$$

where $S_{n_l^m}$ is the set of nodes that have successfully received packet m at time n_l^m due to transmission from node $i_{n_l^m,m}$. We call event B_k^m a mis-routing of order k. For $m > m_0$,

$$Prob(B_k^m) \leq (\max_{i,S} \epsilon_n(i,S))^k \leq \gamma^k.$$

Now for packets $m > m_0$, let us consider the expected differential reward under policies π^* and ϕ^*:

$$\begin{aligned}
&\mathbf{E}^{\pi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}} | \mathcal{F}_0\right\}\right] - \mathbf{E}^{\phi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \\
&= V^*(o) - \mathbf{E}^{\phi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \\
&= \sum_{k=0}^{\infty} \mathbf{E}^{\phi^*}\left[V^*(o) - \left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\} | B_k^m\right] \\
&\quad \times Prob(B_k^m) \\
&\leq \sum_{k=0}^{\infty} k\ R\ Prob(B_k^m) \qquad (20) \\
&\leq R\sum_{k=1}^{\infty} k\gamma^k \qquad (21) \\
&= \delta, \qquad (22)
\end{aligned}$$

where $\delta = \frac{\gamma R}{(1-\gamma)^2}$. Inequality (20) is obtained by noticing that maximum loss in the reward occurs if algorithm d-AdaptOR decides to drop packet m (no reward) while there exists a node j in the set of potential forwarders such that $V^*(j) \approx R$.

Thus, for all $\delta > 0$ the expected average per packet reward under policy ϕ^* is bounded as

$$\begin{aligned}
&\liminf_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m-1} c_{i_{n,m}}\right\}\right] \\
&\geq \liminf_{N\to\infty} E^{\phi^*}\left[\frac{1}{M_N}\sum_{m=1}^{M_N}(V^*(o) - \delta)\right] \\
&= V^*(o) - \delta.
\end{aligned}$$

■

ACKNOWLEDGMENT

The authors would like to thank A. Plymoth and P. Johansson for the valuable discussions.

REFERENCES

[1] C. Lott and D. Teneketzis, "Stochastic Routing in Ad hoc Wireless Networks," *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, vol. 3, pp. 2302–2307 vol.3, 2000.

[2] P. Larsson, "Selection Diversity Forwarding in a Multihop Packet Radio Network with Fading channel and Capture," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 2, no. 4, pp. 4754, October 2001.

[3] M. Zorzi and R. R. Rao, "Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Multihop Performance," *IEEE Transactions on Mobile Computing*, vol. 2, no. 4, 2003.

[4] S. Biswas and R. Morris, "ExOR: Opportunistic Multi-hop Routing for Wireless Networks," *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 3344, October 2005.

[5] S.R. Das S. Jain, "Exploiting Path Diversity in the Link Layer in Wireless Ad hoc Networks," *World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005. Sixth IEEE International Symposium on a*, pp. 22–30, June 2005.

[6] C. Lott and D. Teneketzis, "Stochastic Routing in Ad-hoc Networks," *IEEE Transactions on Automatic Control*, vol. 51, pp. 52–72, January 2006.

[7] E. M. Royer and C.K. Toh, "A Review of Current Routing Protocols for Ad-hoc Mobile Wireless Networks," *IEEE Pers. Communications*, vol. 6, pp. '46–55, April 1999.

[8] T. Javidi and D. Teneketzis, " Sensitivity Analysis for Optimal Routing in Wireless Ad Hoc Networks in Presence of Error in Channel Quality Estimation," *IEEE Transactions on Automatic Control*, pp. '1303–1316, August 2004.

[9] W. Usahaa and J. Barria, "A Reinforcement Learning Ticket-Based Probing Path Discovery Scheme for MANETs ," *Elsevier Ad Hoc Networks*, vol. 2, April 2004.

[10] H. Satoh, "A Nonlinear Approach to Robust Routing Based on Reinforcement Learning with State Space Compression and Adaptive Basis Construction," *IEICE Transactions Fundamentals*, vol. 91-A, January 2008.

[11] Shyamnath Gollakota and Dina Katabi , "ZigZag Decoding: Combating Hidden Terminals in Wireless Networks," in *ACM SIGCOMM*, 2008.

[12] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: John Wiley & Sons, 1994.

[13] Sidney Resnick, *A Probability Path*, Birkhuser, Boston, 1998.

[14] Dimitri P. Bertsekas and John N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[15] William Stallings, *Wireless Communications and Networks*, Prentice Hall, second edition, 2004.

[16] J. Doble, *Introduction to Radio Propagation for Fixed and Mobile Communications* , Artech House, Boston, 1996.

[17] M Kurth, A Zubow, and JP Redlich, "Cooperative Opportunistic Routing Using Transmit Diversity in Wireless Mesh Networks ," in *INFOCOM*, April 2008, pp. 1310–1318.

[18] P. Auer, "Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning ," in *Advances in Neural Information Processing Systems*, 2007.

[19] Parul Gupta and Tara Javidi, " Towards Throughput and Delay Optimal Routing for Wireless Ad-Hoc Networks," in *Asilomar Conference*, November 2007, pp. 249–254.

[20] M. J. Neely, "Optimal Backpressure Routing for Wireless Networks with Multi-Receiver Diversity," in *Conference on Information Sciences and Systems (CISS)*, March 2006.

[21] J.N. Tsitsiklis, "Asynchronous Stochastic Approximation and Q-learning," *Proceedings of the 32nd IEEE Conference on Decision and Control*, vol. 1, pp. '395–400, Dec 1993.

[22] L. Breiman, *Probability*, Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1992.

Abhijeet Bhorkar has received the B.Tech. degree and M.Tech. degree in the electrical engineering from the Indian Institute of Technology, Bombay, both in 2006. He is currently working toward the Ph.D. degree at the University of California, San Diego.

His research interests are primarily in the areas of stochastic control and estimation theory, information theory, and their applications in the optimization of wireless communication systems.

Mohammad Naghshvar received the B.S. degree in electrical engineering from Sharif University of Technology in 2007. He is currently pursuing the M.S./Ph.D. degree in electrical and computer engineering at the University of California, San Diego.

His research interests include stochastic control theory, network optimization, and wireless communication.

Tara Javidi (S'96-M'02) studied electrical engineering at the Sharif University of Technology from 1992 to 1996. She received her MS degrees in Electrical Engineering: Systems, and Applied Mathematics: Stochastics, from the University of Michigan, Ann Arbor, MI. She received her PhD in EECS from University of Michigan, Ann Arbor in 2002.

From 2002 to 2004, Tara was an assistant professor at the electrical engineering department, University of Washington, Seattle. She joined University of California, San Diego, in 2005, where she is currently an assistant professor of electrical and computer engineering. She was a Barbour Scholar during 1999-2000 academic year and received an NSF CAREER Award in 2004. Her research interests are in communication networks, stochastic resource allocation, and wireless communications.

Bhaskar D. Rao (S'80-M'83-SM'91-F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California, San Diego, where he is currently a Professor with the Department of Electrical and Computer Engineering. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human-computer interactions.

Dr. Rao has been a Member of the Statistical Signal and Array Processing Technical Committee of the IEEE Signal Processing Society. He is currently a Member of the Signal Processing Theory and Methods Technical Committee.