An Adaptive Opportunistic Routing Scheme for Wireless Ad-hoc Networks

A.A. Bhorkar, M. Naghshvar, T. Javidi and B.D. Rao
Department of Electrical Engineering,
University of California San Diego, CA, 92093
{abhorkar, naghshvar, tjavidi, brao}@ucsd.edu

Abstract—**In this paper, an adaptive opportunistic routing scheme for multi-hop wireless ad-hoc networks is proposed. The proposed scheme utilizes a reinforcement learning framework to achieve the optimal performance 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 cost 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.

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. Opportunistic routing decisions are made in an on-line manner, choosing the next relay based on the actual transmission outcomes as well as a rank ordering of relays. This on-line and sample-path dependent structure of opportunistic schemes improves the performance of routing by exploiting the broadcast nature of wireless transmissions as well as the inherent path and multi-user diversity present in a network.

The authors in [1], [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] provides a unifying framework for almost all versions of opportunistic routing such as SDF [2], GeRaF [3] and EXOR [4].[1]

The opportunistic algorithms proposed in [1]–[6] implicitly 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". With the exception of [7], which provides a sensitivity analysis of opportunistic routing when channel models are erroneous, by and large, the question of learning and estimating channel statistics has not been explored in the opportunistic routing context. In this paper, using a reinforcement learning framework, we propose an adaptive opportunistic routing (AdaptOR) algorithm which minimizes the expected average per packet cost when zero or erroneous knowledge of transmission success probabilities and network topology is available.

The rest of the paper is organized as follows: In Section II, we discuss the system model and formulate the problem. Section III-A formally introduces our proposed routing algorithm, Adaptive Opportunistic Routing (AdaptOR). We then state the optimality theorem for AdaptOR algorithm in Section III-B. In Section IV, we analyze the convergence and optimality of the algorithm. Finally, we conclude the paper and discuss future work in Section V.

We end this section with a note on the notations used. For a vector $x \in \mathbb{R}^D$, $D \geq 1$, we use $x(l)$ to denote the l^{th} element of the vector. We use n^+ to denote the time just after the start of slot $[n, n+1)$ and $(n+1)^-$ to denote the time just before the end of the slot $[n, n+1)$.

II. SYSTEM MODEL

We consider the problem of routing packets from the source node o to a destination node d in a wireless ad-hoc network of $d+1$ nodes denoted by the set $\Theta = \{o, 1, 2, \cdots, d\}$. The time is slotted and indexed by $n \geq 0$. A packet indexed by $m \geq 0$ is generated at the source node o at time τ_s^m according to an arbitrary distribution with stabilizable rate $\lambda > 0$.

We assume that the successful reception of the packet transmitted by a node occurs according to a fixed conditional probability distribution over the set of nodes in the network. Furthermore, we assume that successful transmissions over different time slots are independent and identically distributed. In particular we characterize the behavior of the wireless channel using a probabilistic *local broadcast model* [6]. The local broadcast model is defined using the transition probability $P(S|i), S \subseteq \Theta, i \in \Theta$, where $P(S|i)$ denotes the probability of successful reception of packet transmitted by node i by all the nodes in S. Note that for all $S \neq S'$, successful reception at S and S' are mutually exclusive and $\sum_{S \subseteq \Theta} P(S|i) = 1$. Logically, node i is always a recipient of

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

[1] 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 is computed so as to minimize the expected number of transmissions (ETX), while GeRaF uses the smallest expected geographical distance from the destination as a criterion for selecting the next-hop.

its own transmission, i.e. $P(S|i) = 0$ if $i \notin S$. Local broadcast model generalizes the notion of link and allows for correlation of successful receptions. When successful transmission to various nodes are independent, $P(S|i)$ can be written as $\Pi_{j \in S} P_{ij}$ where $0 \leq P_{ij} \leq 1$ represents the link quality. The successful reception of the packet by the neighbors is assumed to be known at the centralized controller with zero error and propagation delay.

Given a successful transmission from node i to the set of nodes S, the next (possibly randomized) routing decision includes 1) retransmission by node i, 2) relaying packet by a node $j \in S$, or 3) dropping the packet all together. If the controller decides to use node j for relay, then node j is assumed to transmit the packet at the next slot, while other nodes $k \neq j, k \in S$ drop that packet.

We assume upon a transmission from node i a fixed transmission cost $c_i > 0$ is incurred. 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.

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 at which 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 if the packet is terminated (dropped) before it reaches the destination, no reward is obtained. Let r_m denote the 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.

Given the assumptions and model, the routing scheme can be viewed as selecting a (possibly random) sequence of nodes $\{i_{n,m}\}$ for relaying packets $m = 1, 2, \cdots$. As such, the expected average per packet reward associated with routing packets along sequence of $\{i_{n,m}\}$ is:

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

where M_N denotes the number of packets terminated upto time N, $i_{n,m}$ denotes the index of the node which transmits packet m at time n, and the expectation is taken over the events of transmission decisions, successful packet receptions, and packet generation times.[2]

Problem (P) : We are interested in maximizing (1) by choosing the sequence of relay nodes $\{i_{n,m}\}$ in the absence of knowledge about the local broadcast model.

In proposing a solution to the Problem **(P)**, we will need the following definitions of action space, state space, and reward function associated with each packet m. The set of all actions, action space, is given by,

$$\mathcal{A} = \Theta \cup \{f\},$$

i.e. the set of relay nodes along with the termination action f. The state space is given by a set $\mathfrak{S}$,

$$\mathfrak{S} = \cup_{i \in \Theta} \{S : P(S|i) > 0\} \cup \{F\},$$

denoting the sets of potential reception outcomes from every node $i \in \Theta$ together with a termination state F. The termination state F is the state visited by the system when termination action f is chosen, i.e. $P(F|f) = 1$. Given a set S of nodes that have received a packet from one of the nodes in Θ, the set of allowable actions is denoted by $A(S) = S \cup \{f\}$. The allowable action in the termination state F is f, i.e. $A(F) = \{f\}$. Without loss of generality, the allowable action associated with any set $S \in Z_d = \{S : d \in S, S \in \mathfrak{S}\}$ is restricted to f, i.e. $A(S) = \{f\}$.

It remains to define the reward function $g : \mathfrak{S} \times \mathcal{A} \to \mathbb{R}$ to represent the reward obtained from taking an action at a given state. In summary, $g(S, a)$ is given as:

$$g(S,a) = \begin{cases} -c_i & a = i \in S \\ R & a = f \ , \ S \in Z_d \\ 0 & a = f \ , \ S \notin Z_d \end{cases} .$$

Let $S_{n,m}$ and $a_{n,m}$ be respectively the state of the system and the routing decision at time n for packet m. Let admissible routing policy ϕ be a sequence of actions $\{a_{\tau_s^m,m}, a_{\tau_s^m+1,m}, \cdots\}$ for all packets m taking values on the allowable action space $A(S)$. For $a \in A(S)$, the event $\{a_{n,m} = a\}$ belongs to the σ-field $\mathcal{H}_n$ generated by $\cup_m \{\tau_s^m, S_{\tau_s^m,m}, a_{\tau_s^m,m}, \cdots, S_{n-1,m}, a_{n-1,m}, S_{n,m}\}$ for all m such that $\tau_s^m \leq n$. Furthermore, let Φ denote the set of admissible policies for Problem **(P)**.

III. THE ALGORITHM AND MAIN RESULTS

A. *Algorithm AdaptOR*

In this section, we present an Adaptive Opportunistic Routing (AdaptOR) algorithm to solve Problem **(P)**. At each time slot n, the algorithm uses a score vector Λ_n in $\mathbb{R}^v$, where $v = \sum_{S \in \mathfrak{S}} A(S)$ is the cardinality of the domain $\mathfrak{S} \times \mathcal{A}$.

Remark $\Lambda_n(S, a)$ evaluated at state $S \in \mathfrak{S}$ and action $a \in A(S)$, can be considered to be an estimate of the expected reward obtained by taking action a at state S at time slot n.

AdaptOR is parametrized by a scaler constant $0 < \gamma \leq 1$ and a sequence of positive scalers $\{\alpha_n\}_{n=1}^{\infty}$. During any time slot [n,n+1), the algorithm uses two counting random variables $\nu_n(S, a)$, $N_n(S)$, and two random sets W_n and Y_n to update the n^{th} iterate Λ_n. Counting random variables $\nu_n(S, a)$ and $N_n(S)$ are equal to the number of times state-action pair (S, a) and state S have been reached upto time n, respectively. Random set $W_n \subseteq \Theta$ denotes the set of transmitting nodes during time slot $[n - 1, n)$, while random set Y_n consists of the set of potential relays associated with transmissions from nodes in W_{n-1}.

---

[2] Our main result establishes the existence of an optimal policy which maximizes the lim in (1) 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 [8, Page 344].

Random counters ν_n, N_n, random sets Y_n, W_n, and Λ_n are initialized as follows:

$$\nu_0(S,a) = 0, N_0(S) = 0,$$

$$Y_0 = \{o\}, W_0 = \{o\},$$

$$\Lambda_0(S,a) = \begin{cases} -R & \text{if } (S,a) = (F,f) \\ 0 & \text{otherwise} \end{cases} .$$

To better conceptualize the working of algorithm AdaptOR, we divide the execution of the algorithm into three stages of reception, adaptive computation, and relay/transmission.

1) **Reception and Acknowledgment Stage**:
This stage is assumed to occur at time n. $W_n \subseteq \Theta$ denotes the (random) set of nodes each of which has transmitted one packet at time n^-. For any transmitter node $a \in W_n$, let S_n^a denote the (random) set of nodes that have successfully received the packet from node a. In the reception and acknowledgment stage the successful reception of the transmitted packet is acknowledged by all the nodes in the set S_n^a for all $a \in W_n$. These nodes form the set of potential relays for node a; collectively they form random set Y_{n+1}, i.e.

$$Y_{n+1} := \{S_n^a : \forall a \in W_n\}.$$

Upon reception and acknowledgment, the counting random variables are incremented as follows:

$$N_n(S) = \begin{cases} N_{n-1}(S) + 1 & \text{if } S \in Y_{n+1} \\ N_{n-1}(S) & \text{if } S \notin Y_{n+1} \end{cases} ,$$

and

$$\nu_n(S,a) = \begin{cases} \nu_{n-1}(S,a) + 1 & \text{if } (S,a) \in Y_n \times W_n \\ \nu_{n-1}(S,a) & \text{if } (S,a) \notin Y_n \times W_n \end{cases} .$$

2) **Adaptive Computation Stage**:
This stage is assumed to occur at n^+. In this stage, for all $(S,a) \in Y_n \times W_n$, Λ_n is updated as follows:

$$\begin{aligned} \Lambda_n(S,a) = \Lambda_{n-1}(S,a) + \\ \alpha_{\nu_n(S,a)} \Big(- \Lambda_{n-1}(S,a) + g(S,a) \\ + \max_{j \in A(S_n^a)} \Lambda_{n-1}(S_n^a, j) \Big) . \end{aligned} \tag{2}$$

For the state-action pair $(S,a) \notin Y_n \times W_n$, Λ_n remains unchanged as

$$\Lambda_n(S,a) = \Lambda_{n-1}(S,a).$$

3) **Relay/Transmission Stage:**
This stage is assumed to occur at $(n+1)^-$. In this stage, the next set of relay nodes (actions) are selected. In particular, for all $S \in Y_{n+1}$, random action $a_{n+1}^S \in A(S)$ is selected according to the following (randomized) rule:

- with probability $(1 - \epsilon_n(S))$,

$$a_{n+1}^S \in \arg\max_{j \in A(S)} \Lambda_n(S,j)$$

is selected,[3] and

- with probability $\frac{\epsilon_n(S)}{|A(S)|}$, $a_{n+1}^S \in A(S)$ is selected randomly, where

$$\epsilon_n(S) = \frac{\gamma}{N_n(S) + 1} .$$

At time $(n+1)^-$, the set of transmitters $W_{n+1} = \{a : \forall S \in Y_{n+1}, a \in \Theta \text{ and } a = a_{n+1}^S\}$ is updated.
All nodes in W_{n+1} transmit a packet at time $(n+1)^-$.

B. *Optimality of AdaptOR*

We will now state our main result on the optimality of AdaptOR, $\phi^* \in \Phi$. Theorem 1 below shows that the expected reward obtained by ϕ^*(AdaptOR) maximizes (1).

Theorem 1. *For all $\phi \in \Phi$,*

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

IV. PROOF

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

A. *Convergence of Λ_n*

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

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

Let $\Lambda^* \in \mathbb{R}^v$ denote the fixed point of operator U,[4] i.e.

$$\Lambda^*(S,a) = g(S,a) + \sum_{S'} P(S'|a) \max_{j \in A(S')} \Lambda^*(S',j), \tag{3}$$

$$\Lambda^*(F,f) = -R. \tag{4}$$

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

Theorem 2. *Let*
(J1) $\Lambda_0(.,.) = 0$ *and* $\Lambda_0(F,f) = -R$,
(J2) $\sum_{l=0}^{\infty} \alpha_l = \infty$, $\sum_{l=0}^{\infty} \alpha_l^2 < \infty$.
Then iterate Λ_n obtained by the stochastic recursion (2) converges to Λ^* almost surely.

Proof: The proof follows using known results on the convergence of a certain super martingale process presented in Theorems 1, 2 in [10]. The detailed proof is provided in [9]. ∎

---

[3] In case ambiguity, node with smallest index is chosen.
[4] Existence and uniqueness of Λ^* is provided in [9].

B. Proof of optimality

Using the convergence result of Λ_n, next we show that the expected average per packet reward under 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 AdaptOR algorithm for Problem **(P)**, we take cue from known results of a closely related Auxiliary Problem **(AP)** wherein the controller has perfect knowledge of local broadcast model as presented in [1], [6].

Let $\mathcal{F}_n$ be the product σ-field $\mathcal{P} \times \mathcal{H}_n$ [11], where $\mathcal{P}$ is the borel σ-field generated by the random probability measures for the local broadcast model.[5] For Auxiliary Problem **(AP)**, let admissible routing policy π be a sequence of actions $\{a_{\tau_s^m,m}, a_{\tau_s^m+1,m}, \cdots\}$ for packet m taking values on the allowable action space $A(S)$ such that the event $\{a_{n,m} = a\}$ belongs to the σ-field $\mathcal{F}_n$. Furthermore, 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

$$J^{\pi}(\{o\}) := \mathbf{E}^{\pi} \left[\left\{ r_m - \sum_{n=0}^{\tau_e^m - 1} c_{i_{n,m}} \right\} | \mathcal{F}_0 \right], \tag{5}$$

where $\mathcal{F}_0 = \mathcal{P}$, and the expectation E^{π} is taken with respect to the random events as well as the conditional distributions over action space defined by policy π. 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,

$$J^{\pi^*}(\{o\}) = \sup_{\pi \in \Pi} J^{\pi}(\{o\}). \tag{6}$$

Auxiliary Problem **(AP)** has been extensively studied in [1], [6], [12] and the following theorem is established in [6].

Fact 1 (Theorem 2.1 [6]). There exists a function $\pi^* : \mathfrak{S} \to \mathcal{A}$ such that the policy $a_{n,m} = \pi^*(S_{n,m})$ is an optimal solution for the Auxiliary Problem **(AP)**.[6] Furthermore, π^* is such that

$$\pi^*(S) \in \arg\max_{j \in A(S)} V^*(j), \tag{7}$$

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

$$\begin{aligned} V^*(d) &= R & (8)\\ V^*(i) &= \max(\{-c_i + \sum_{S'} P(S'|i)(\max_{j \in S'} V^*(j))\}, 0) & (9)\\ V^*(f) &= 0. & (10) \end{aligned}$$

Lastly, $V^*(j)$ is the maximum expected reward for routing a packet from node j to destination d:

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

Lemma 1 below states the relationship between the solution of Problem **(P)** and that of the Auxiliary Problem **(AP)**. More specifically, Lemma 1 shows that $V^*(o)$ is an upper bound for the solution to Problem **(P)**.

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

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

Proof: The proof is given in [9]. Intuitively the result holds because the set of admissible policies Φ in **(P)** is a subset of admissible policies Π in **(AP)**. ■

Lemma 2 gives the achievability proof for Problem (P) by showing that the expected average per packet reward of AdaptOR is no less than $V^*(o)$.

Lemma 2. *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: The proof is given in Appendix A. ■

Lemmas 1 and 2 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 $V^*(o)$. This together with Lemma 1 establishes the proof of Theorem 1.

V. CONCLUSIONS

In this paper, we proposed an adaptive opportunistic routing scheme which maximizes the expected average per packet reward from the source to the destination in absence of knowledge regarding network topology and link qualities.

We would like to point out that AdaptOR can be readily extended to scenarios in which the routing decisions and computations are done in a decentralized and asynchronous manner. We refer interested readers to [9].

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 arises in many TDM based schemes such as [13]. The joint design of MAC and routing remains an important area of future research.

---

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

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

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] 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.

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

[9] A.A. Bhorkar, M. Naghshvar, T. Javidi, and B.D. Rao, "A routing policy for wireless ad-hoc networks," Tech. Rep., University of California, San Diego, 2008, Available at http://dsp.ucsd.edu/~bhorkar/files/Adaptor_journal.pdf.

[10] 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.

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

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

[13] M.J. Neely, E. Modiano, and C.E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *INFOCOM 2003*, vol. 1, pp. 745–755 vol.1, March 2003.

APPENDIX

A. Proof of Lemma 2

Proof: From (3), (4), (8), (9) we obtain the following equality

$$\arg\max_{j \in A(S)} V^*(j) \quad = \quad \arg\max_{j \in A(S)} \Lambda^*(S,j). \tag{11}$$

Let

$$b = \min_{S \in \mathfrak{S}} \min_{\substack{i,j \in A(S) \\ \Lambda^*(S,i) \neq \Lambda^*(S,j)}} \frac{|\Lambda^*(S,i) - \Lambda^*(S,j)|}{2}. \tag{12}$$

Theorem 2 implies that, in an almost sure sense, there exists packet index $m_1 < \infty$ such that for all $n > \tau_s^{m_1}$,

$$|\Lambda_n(S,a) - \Lambda^*(S,a)| \leq b \quad \forall S \in \mathfrak{S}, a \in A(S). \tag{13}$$

Therefore, from time $\tau_s^{m_1}$ onwards, given any set of S, probability that algorithm AdaptOR chooses an action $a \in A(S)$ such that $\Lambda^*(S,a) \neq \max_{j \in A(S)} \Lambda^*(S,j)$ is upper bounded by $\epsilon_n(S)$. Furthermore, since each state is visited infinitely often [9] ($N_n(S) \to \infty$) there exists packet index $m_2 < \infty$ almost surely such that for all $n > \tau_s^{m_2}$, $\max_S \epsilon_n(S) < \delta$ for a given $\delta > 0$.

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 $m_0 R_{max} < \infty$ and lower-bounded by $-\frac{m_0}{\lambda} d \max_i c_i > -\infty$, hence their presence does not impact the expected average reward. Consequently, we only need to consider the errors due to random decisions of policy ϕ^* (exploration) 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 at which routing algorithm routes packet m differently from the possible set of optimal actions. Mathematically speaking, event B_k^m occurs iff there exists instances $\tau_s^m \leq n_1^m \leq n_2^m \cdots n_k^m \leq \tau_e^m$ such that for all $l = 1, 2, \cdots, k$

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

where $S_{n_l^m}$ is the set of nodes which have successfully received packet m at time n_l^m. We call such events B_k^m a mis-routing of order k. It is straight-forward to show that for $m > m_0$,

$$Prob(B_k^m) \leq \delta^k.$$

For any packet m, $m > m_0$, let us consider the expected differential reward under policies π^* and ϕ^*:

$$\begin{aligned}
&\mathbf{E}^{\pi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}} | \mathcal{F}_0\right\}\right] - \mathbf{E}^{\phi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}}\right\}\right] \\
&= \quad V^*(o) - \mathbf{E}^{\phi^*}\left[\left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}}\right\}\right] \\
&= \quad \sum_{k=0}^{\infty} \mathbf{E}^{\phi^*}\left[V^*(o) - \left\{r_m - \sum_{n=\tau_s^m}^{\tau_e^m - 1} c_{i_{n,m}}\right\} | B_k^m\right] \\
&\quad\quad \times Prob(B_k^m) \\
&\leq \quad \sum_{k=0}^{\infty} k \; R \; Prob(B_k^m) \qquad (14) \\
&\leq \quad R \sum_{k=1}^{\infty} k \delta^k \qquad (15) \\
&= \quad \delta', \qquad (16)
\end{aligned}$$

where $\delta' = \frac{\delta R}{(1-\delta)^2}$. Inequality (14) is obtained by noticing that maximum loss in the reward occurs if algorithm 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 the expected average per packet reward under policy ϕ^* is bounded as

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

■