Cooperative Transmission Scheme of Energy Harvesting Tags

Mengistu Abera Mulatu¹*, Wiseman Nyembe¹, Yeneneh Tamirat Negash²

¹Department of Electrical and Electronic Engineering, University of Swaziland, Kwaluseni, Swaziland
²Department of Business Administration, Asia University, Taichung, Taiwan

Email address:
mamulatu@uniswa.sz (M. A. Mulatu), wnyembe@uniswa.sz (W. Nyembe), yeneneh@asia.edu.tw (Y. T. Negash)

Corresponding Author

To cite this article:
Mengistu Abera Mulatu, Wiseman Nyembe, Yeneneh Tamirat Negash. Cooperative Transmission Scheme of Energy Harvesting Tags. International Journal of Wireless Communications and Mobile Computing. Vol. 5, No. 5, 2017, pp. 26-31.
doi: 10.11648/j.wcmc.20170505.11

Abstract: Energy harvesting tags with cooperative communication capabilities are recently emerging as a viable infrastructure for internet of things (IoT) applications. This letter studies the cooperative transmission strategy for a network of energy harvesting active networked tags (EnHANTs), that is adapted to the available energy resource and identification request. We consider a network of tags to communicate with the reader either directly or by cooperating with neighboring tags. We formulate the problem as a Markov decision process (MDP). The simulation results are provided to show the performance of the cooperative transmission policy under various energy harvesting scenarios.

Keywords: Cooperative Communication, Energy Harvesting Tags, Markov Decision Process, Transmission Scheme

1. Introduction

Cooperative communication and energy harvesting are practical solutions to overcome the battery and communication reliability problems in wireless devices. Energy-harvesting active networked tags (EnHANTs) are recently proposed as tiny devices that can be attached to common place objects [1], [2]. EnHANTs can also be applied to tracking and monitoring. They can communicate with one another and with EnHANT-friendly devices to cooperate and forward information to the intended destination. In order to effectively utilize the random energy resource and maintain reliability, efficient cooperative transmission scheduling needs to be designed. Wang et al. [3] proposed an optimal transmission policy for single link system. An EnHANT-equipped object might lack sufficient energy to respond directly to the reader when the reader is outside the communication range. In such a case, to sustain communication, the object can link its information to a neighboring object, in which the neighboring object forwards the information to the reader by using the concept of a relay [4] -- [6]. We consider an amplify and forward (AF) relaying because of its low level of complexity. In an AF, the relay simply amplifies and forwards the received data to the destination. In [7], the use of energy-harvesting nodes as AF cooperative relays that assist communication of source and destination was proposed. In [8], the authors studied energy efficient scheduling strategies for wireless sensor networks with energy-harvesting. They considered a case where a node may use either direct transmission or a cooperative relay transmission and formulated the problem using Markov decision process (MDP).

In this letter, we consider a network of EnHANTs in which a tag has two options (i.e. direct and cooperative) to communicate with the reader. The relay tag assists its neighbor without affecting its own transmission. We assume that energy detection technique and accurate synchronization timer is employed at each tag. We consider a cooperative transmission strategy that optimizes the long-term average throughput by taking into account both the identification request state and energy constrains.

2. System Model

A. Communication System Model
We consider a network of three tags (T_1, T_2, T_3) and a reader (R), in which each tag communicates with a reader by using the direct and the cooperative modes. Assume that AF relaying protocol is employed and all copies of the relayed signal of a given tag are combined using maximum ratio combining (MRC) [7] at the reader side so as to achieve diversity gains. Each tag can directly communicate with R if it has sufficient battery energy or it can communicate with the assistance of a neighboring tag if its energy cannot support direct communication. The neighboring tag cooperates only if its stored energy is sufficient to receive and relay the data.

A tag can have different battery states distinguished by three battery thresholds (δ_1^T, δ_2^T, δ_3^T, i=1, 2, 3.) If the battery energy level is below the minimum threshold, δ_0^T, T_i cannot respond to reader requests. This means that the battery energy is either empty or below this minimum value which is not enough for transmission of data. If the battery is above δ_1^T and below a second threshold δ_2^T, then T_i can respond to the reader’s request only through the assistance of a neighboring tag since T_i is out of reach of the reader. On the contrary, if the battery energy level is above δ_1^T and below a third threshold δ_3^T, T_i can independently communicate with the reader upon request. However, a tag whose battery is in this state has no relaying capability for neighboring tags. This is because it requires additional energy to receive the data from the neighboring tag and forward them to the reader. Unconventionally, transmitting is cheaper than receiving in terms of energy in EnHANTs. If a tag has a battery energy level above δ_3^T, then it can relay data for a neighboring tag as well as it can communicate its own data with the reader.

It is assumed that communication takes place in time slotted fashion and battery energy parameter exchanges are conducted between tags and a reader before any transmission occurs. During the first phase of T_i’s time slot, T_i transmits packets that contain information symbols to the reader. Without loss of generality, each transmitted packet from T_i of sufficient energy can assist relaying T_j’s information. The cooperative mode occurs during two equal transmission phases of T_i’s time slot. If the battery energy level is below the minimum threshold, the signal to T_j and R due to the transmitted information from T_i in the first phase are represented as γ_{T_i}^R = δ_1^T \gamma_{T_i}^T + n_{R_i}^T and γ_{T_j}^R = δ_1^T \gamma_{T_j}^T + n_{R_j}^T, respectively, where δ_{T_i,T_j} and δ_{T_i,R} are the channel coefficients from T_i to T_j and R, and n_{T_i}^T and n_{T_j}^T are the ambient Gaussian noises at T_i and R with zero mean and equal variance N_0. T_i amplifies the information and forwards it to the reader. Accordingly, the received signal vector at R from T_i in the second phase is γ_{T_i}^R = \Gamma \theta_{T_i,R} \gamma_{T_i}^T + n_{R_i}^T, where γ is the signal amplification factor of T_i and written as γ^2 = (\Gamma^2 + (\theta_{T_i,R}^2 + N_0)), θ_{T_i,R} is the channel coefficient from T_i to R, n_{R_i}^T is the ambient Gaussian noise and ε_{T_i}^T is the transmitted signal energy of T_i. We assume that the reader decodes the information after combining the signals received from T_i and T_j by using MRC. Accordingly, the total end-to-end SNR at R when T_i is used as an AF relay can be written as [9], [10]:

\[ γ_{T_i,T_j,R} = γ_{T_i,R} + \frac{γ_{T_i}γ_{T_i}R}{1+γ_{T_j}R + γ_{T_i}R} \]  

(1)

where γ_{T_i}T_j = \delta_2^T \gamma_{T_i}^T x_{T_j} / N_0 is the instantaneous SNR of T_i at T_j and γ_{T_j,R} = \delta_2^T \gamma_{T_j}^T / N_0 is the instantaneous SNR of T_j at R. Therefore, the SNR at the reader from T_i can be written as follows using either of the modes:

\[ γ_{T_i} = \begin{cases} γ_{T_i,R} \quad \text{direct mode} \\ γ_{T_i,T_j,R} \quad \text{cooperative mode} \end{cases} \]  

(2)

We consider the path loss effect based on the tags’ position. Let d_{T_i,R} and d_{T_i,T_j} ≠ j be the distances of the T_i, R and T_i, T_j links, respectively. Without loss of generality, we can model θ_{T_i,R} = d_{T_i,R}^{-η} and θ_{T_i,T_j} = d_{T_i,T_j}^{-η}, where η is the path loss exponent.

We assume that the reader processes the received PPM signal by using a compressive sensing (CS) technique and signal detection methods [3], [11] to avoid the need to employ high sampling rate A/D converters. When using this detection method, the probability of mis-detecting a PPM symbol is expressed as follows [3]:

\[ p_{md} = 1 - Q \left( -\frac{M}{\sqrt{N}} γ_{T_i} \right)^{K-1} \]  

(3)

Where \( Q(x) = \frac{1}{\sqrt{2π} e^{-x^2}} \int_{-\infty}^{x} e^{-t^2} dt \), and γ_{T_i} is determined using (2). The mis-detection error (3) occurs when the reader fails to successfully decode the data transmitted by a tag. In addition, the reader may not get any response when a tag lacks energy in its battery, leading to the probability...
of no-response error, \( p_{n_{T}} = \frac{1}{e^{x_{k}(T_{i}=0)}} \), where \( e_{k}(T_{i}) \) is the symbol energy of \( T_{i} \) at the \( k \)th time interval and the indicator function \( 1_{\{x_{k}\}} \) is defined as \( 1_{\{x_{k}\}} = 1 \) if \( X \) is true, and 0 otherwise. Accordingly, at any given time interval \( k \), the

\[
p_{e_{k}}(e_{k}) = (\beta \cdot p_{n_{T}}(e_{k}) + (1 + \beta)\rho_{m_{T}}(e_{k} \cdot 1_{\{e_{k}>0\}}) \cdot \alpha_{k}
\]

where \( \beta \in [0, 1] \) is a weighting factor that constitutes the two errors under one performance metric. \( \alpha_{k} = 1 \) indicates a reader requests and \( \alpha_{k} = 0 \) indicates no request.

**B. Energy Harvesting Dynamics**

We assume that each tag has a finite rechargeable battery capacity and light harvesting device. The reader is assumed to have no power constraint. Let \( B_{k}^{T_{i}} \) be the battery energy of \( T_{i} \) at the beginning of the \( k \)th time interval and \( h_{k}^{T_{i}} \) be the amount of energy harvested by \( T_{i} \) during the \( k \)th time interval.

We consider the random energy arrival process and assume that \( h_{k}^{T_{i}} \) takes discrete values from the set \( H = \{ H_{1}, H_{2}, ..., H_{D} \} \). Let \( q_{k}^{n_{H_{k}+1}} \) be the state transition probability from state \( h_{k} \) to \( h_{k+1} \). Denote \( q_{k}^{H_{1}}, q_{k}^{H_{2}}, ..., q_{k}^{H_{D}} \) to be the steady state energy harvesting probability corresponding to \( H_{1}, H_{2}, ..., H_{D} \), respectively. The request of the reader at the beginning of the \( k \)th time interval is modeled as \( a_{k} \sim Bernoulli (r) \). The energy stored in the battery of \( T_{i} \) at time interval \( k \) for use in the subsequent time interval is determined as:

\[
B_{k+1}^{T_{i}} = \min (B_{k}^{T_{i}} - \alpha_{k} W_{k}^{T_{i}} + h_{k}^{T_{i}}, B_{k}^{T_{i}}_{\text{max}})
\]

where \( W_{k}^{T_{i}} = L e_{k}^{T_{i}} + b_{k}^{T_{i}} \), \( C_{k}^{T_{i}} \) and \( B_{k}^{T_{i}}_{\text{max}} \) is the maximum battery capacity. \( h_{k}^{T_{i}} = 1 \) if \( T_{i} \) cooperates, and 0 otherwise. \( C_{k}^{T_{i}} \) includes energy to exchange state information before data transmission attempt and energy to transmit and receive

\[
U(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \left( \nu_{T_{i}}(e_{k}) + \nu_{T_{j}}(e_{k}) \right) \nu_{T_{s}}(e_{k})
\]

and the optimization problem that maximizes the average packet throughput can be formulated as follows:

\[
\mathcal{U}: \max_{\pi \in \mathcal{A}} \mathcal{U}(\pi)
\]

This optimization problem can be computed using an MDP to obtain an optimal policy.

**B. Markov Decision Process Formulation**

An MDP is defined using a quadruplet \((S,A,p_{a}(s_{i}, s_{j}), R_{a}(s_{i}, s_{j}))\), where \( S \) is the set of states, \( A \) is the set of actions, \( p_{a}(s_{i}, s_{j}) \) denotes the transition probability from state \( s_{i} \) to state \( s_{j} \) when an action \( a_{i} \in A \) is taken, and \( R_{a}(s_{i}, s_{j}) \) is the reward due to transitions from state \( s_{i} \) to state \( s_{j} \) when an action \( a_{i} \) is taken. The goal of an MDP is to choose a policy \( \pi \) that assigns actions to each state and maximizes the average reward. At any time interval \( k \), \( T_{1}, T_{2} \) and \( T_{3} \) consume the joint energy \( A_{i} \) to send their packets to the reader either directly or by cooperating. The joint choice \( A_{i} \) causes a state change from \( S_{i} \) to \( S_{e_{i+1}} \), yielding the immediate reward of \( R_{A_{i}}(S_{i}, S_{e_{i+1}}) \cdot R_{A_{i}}(S_{i}, S_{e_{i+1}}) \) is weighted of these two error probabilities results in communication error probability in EnhANTs is defined as follows [3]:

\[
\nu_{T_{i}}(e_{k}) = \left( 1 - p_{e_{k}}(e_{k}) \right) \cdot \alpha_{k}; \text{ direct mode}
\]

\[
\nu_{T_{i}}(e_{k}) = \left( \frac{1}{2} \right) \cdot \alpha_{k}; \text{ co-op mode}
\]

\[
\nu_{T_{i}}(e_{k}) = \left( 1 - p_{e_{k}}(e_{k}) \right) \cdot \alpha_{k}; \text{ direct mode}
\]

\[
\nu_{T_{i}}(e_{k}) = \left( \frac{1}{2} \right) \cdot \alpha_{k}; \text{ co-op mode}
\]

\[
R_{A_{i}}(S_{i}, S_{e_{i+1}}) = \nu_{T_{i}}(e_{k}) + \nu_{T_{j}}(e_{k}) + \nu_{T_{s}}(e_{k})
\]

\[
R_{ib}(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} R_{A_{i}}(S_{i}, S_{e_{i+1}})
\]

3. Problem Formulation

**A. Performance Measure**

Each time slot in the time interval \( k \) is assumed to be able to transmit a packet. Denote the state of the tags as \( S_{k} = (B_{k}, B_{k}^{T_{1}}, B_{k}^{T_{2}}, B_{k}^{T_{3}}, h_{k}^{T_{1}}, h_{k}^{T_{2}}, h_{k}^{T_{3}}, \alpha_{k}) \) and \( S \) be the set of all possible states. The transmission policy \( \pi \) is a mapping from the states \( S \) to the energy consumption \( A \). Given the current state \( S_{k} \) and the policy \( \pi : S \to A \), the packet throughput of tag \( T_{i} \) at the \( k \)th time interval can be expressed as

\[
U_{T_{i}}(e_{k}) = R(1 - p_{e_{k}}(e_{k})) \cdot \alpha_{k}; \text{ direct mode}
\]

\[
R_{i} \cdot \rho_{m_{T}}(e_{k} \cdot 1_{\{e_{k}>0\}}) \cdot \alpha_{k}; \text{ co-op mode}
\]

chosen to equal the throughput and \( R_{i} \) is normalized to one. \( \nu_{T_{i}} \) denotes such normalized throughput corresponding to \( T_{i} \) and can be expressed as follows:

Thus, the infinite horizon average reward of the tags is expressed as:

\[
R_{ib}(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} R_{A_{i}}(S_{i}, S_{e_{i+1}})
\]

\[
A_{i} \cdot (W_{k}^{T_{1}}, W_{k}^{T_{2}}, W_{k}^{T_{3}}) \) be the joint energy consumption of \( T_{1}, T_{2} \) and \( T_{3} \) expressed in finite discrete values and \( \mathcal{A} \) be the set of all possible joint energy consumption by all tags (i.e. \( \mathcal{A} = \{0, a_{1}, a_{2}, ..., a_{f}\} \)). The energy harvesting probability corresponding to \( T_{i} \) takes discrete values from the set \( \mathcal{H} = \{ h_{1}, h_{2}, ..., h_{D} \} \). Let \( q_{k}^{H_{1}}, q_{k}^{H_{2}}, ..., q_{k}^{H_{D}} \) to be the steady state energy harvesting probability corresponding to \( H_{1}, H_{2}, ..., H_{D} \), respectively. The request of the reader at the beginning of the \( k \)th time interval is modeled as \( a_{k} \sim Bernoulli (r) \). The energy stored in the battery of \( T_{i} \) at time interval \( k \) for use in the subsequent time interval is determined as:

\[
B_{k+1}^{T_{i}} = \min (B_{k}^{T_{i}} - \alpha_{k} W_{k}^{T_{i}} + h_{k}^{T_{i}}, B_{k}^{T_{i}}_{\text{max}})
\]

where \( W_{k}^{T_{i}} = L e_{k}^{T_{i}} + b_{k}^{T_{i}} \), \( C_{k}^{T_{i}} \) and \( B_{k}^{T_{i}}_{\text{max}} \) is the maximum battery capacity. \( h_{k}^{T_{i}} = 1 \) if \( T_{i} \) cooperates, and 0 otherwise. \( C_{k}^{T_{i}} \) includes energy to exchange state information before data transmission attempt and energy to transmit and receive

\[
U(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \left( \nu_{T_{i}}(e_{k}) + \nu_{T_{j}}(e_{k}) \right) \nu_{T_{s}}(e_{k})
\]

and the optimization problem that maximizes the average packet throughput can be formulated as follows:

\[
U: \max_{\pi \in \mathcal{A}} \mathcal{U}(\pi)
\]

This optimization problem can be computed using an MDP to obtain an optimal policy.
as follows by using the MDP problem:

$$
\hat{U} : \max_{\pi} \mathbb{E}_{\pi} R_{th}(\pi)
$$

subject to: battery state (5)

(12)

Because all components of $S_k$ are discrete, a finite number of states exists. When the action $A_i$ is taken, the state $S_i$ yields a transition to any of the next possible states $S_{i+1}$. The state transition probability $p_{A_k}(S_k, S_{k+1})$ is the probability that the system will go to state $S_{k+1}$ when action $A_k$ is taken at state $S_k$ during the $k^{th}$ time interval. Since the tags harvest energy independently and the reader randomly requests for information, the state transition probability can be defined as:

$$p_{A_k}(S_k, S_{k+1}) = \sum_{S_{k+1} \in S} \pi_{A_k}(S_k, S_{k+1}) p(S_k, S_{k+1}), \quad (13)$$

The relative value iteration (RVI) algorithm [12] can be applied to compute the optimal cooperative transmission policy for (14).

4. Numerical Results

We evaluate the performance of the optimal cooperative transmission policy by considering various energy harvesting scenarios. We assume that each message transmitted from the tags comprises $L = 4$ PPM symbols with a symbol modulation order of $K = 32$. The symbol energy of each tag is chosen from $T = \{1, 2, 4\}$. The compression ratio $M/N = 0.1$. The battery thresholds $\delta_1, \delta_2, \delta_3$ of the tags are assumed to be 4, 8, and 28 respectively for batteries that exhibit a capacity of $B_{\text{max}} = 40$. Furthermore $\beta = 0.5, \eta = 2, \mathcal{H} = \{0.5, 10\}$ and $C_{T_i} = 12$ are set, assuming that the energy required to receive a symbol is five times more compared with the energy required to transmit a symbol. Moreover, $d_{T_i} = 1, d_{T_j} = 0.5, i, j = 1, 2, 3, i \neq j$. The number of simulated time slots is $N = 10^6$ and the convergence threshold of the RVI algorithm is $\epsilon = 10^{-4}$. Let the direct transmission of a symbol of unit energy yields an SNR of $\frac{\delta_1}{\delta_0} = 4$ dB at the reader.

Figure 1. Performance comparisons under energy balanced scenario.
For performance comparison purpose, we consider the optimal direct transmission policy [3] and cooperative transmission policy for two tags [5]. Figure 1 provides comparisons of the average throughput performance of a tag when the steady-state probabilities of \((H_1, H_2, H_3)\) for \(T_1\), \(T_2\), and \(T_3\) are \((0.33, 0.33, 0.33)\). All the tags harvest equal...
proportion of energy from the environment and can assist each other in forwarding data in this energy balanced scenario. In this scenario, the optimal cooperative policies for both two-tag and three-tag cases achieve equal average throughput performances per tag due to identical energy harvesting environment for all the tags. These performances are superior compared with performance of the non-cooperative policies. Figure 2 shows second scenario when the steady-state probabilities of \((H_1, H_2, H_3)\) for \(T_1, T_2,\) and \(T_3\) are \((0.33, 0.33, 0.33), (0.28, 0.44, 0.28)\) and \((0.11, 0.39, 0.50)\), respectively. \(T_3\) has a better energy harvesting condition and assists both \(T_1\) and \(T_2\). The average throughput per tag of \(\{T_3\}\) outperforms the performance of both \(\{T_1, T_2, T_3\}\) and \(\{T_1, T_2\}\) mainly due to variation in the energy harvesting conditions of the tags. The performance of \(\{T_1, T_2\}\) is inferior compared to others due to worse harvesting conditions of \(T_1\) and \(T_2\). In all cases, the cooperative transmission policy outperforms the direct transmission policy proposed by [3]. Figure 3 depicts the average throughput performance when the steady-state probabilities of \((H_1, H_2, H_3)\) for \(T_1, T_2,\) and \(T_3\) are \(0.11, 0.39\) and \(0.50\).

In this energy surplus scenario, all the tags are at good energy harvesting conditions and tags mainly communicate with the reader independently without helping each other. As a result, the average throughput performance per tag for the proposed three tags and two tags [5] are identical. The optimal direct policy is also close to the proposed policy due to good energy harvesting conditions of the tags. Figure 1 and Figure 3 show that the average throughput performance per tag of the cooperative transmission policy we proposed are identical for two different sizes of tags under the same energy harvesting conditions.

5. Conclusion

In this letter, we formulated the optimal cooperative transmission problem by considering the case of three tags, aiming to maximize the long-term average throughput of tags. We used MDP RVI algorithm to obtain numerical results under different energy harvesting scenarios. The proposed cooperative transmission policy showed better performance in terms of average throughput compared to the direct transmission policy. However, the joint cooperative performance of the three tags and two tags in terms of the average throughput per tag are identical under the same energy harvesting conditions. The results demonstrate that EnHANTS can jointly use the time varying energy resources efficiently and achieve improved communication reliability.

References

[1] M. Gorlatova, P. Kinget, I. Kymissis, D. Rubenstein, X. Wang, and G. Zussman, “Challenge: ultra-low-power energy-harvesting active networked tags (enhants),” in Proc. 2009 ACM Mobile Telecommun. Conf., pp. 253–260.

[2] M. Gorlatova, P. Kinget, I. Kymissis, D. Rubenstein, X. Wang, and G. Zussman, “Energy harvesting active networked tags (enhants) for ubiquitous object networking,” IEEE Trans. Wireless Commun., vol. 17, no. 6, pp. 18–25, Dec. 2010.

[3] Z. Wang, A. Tajer, and X. Wang, “Communication of energy harvesting tags,” IEEE Trans. Commun., vol. 60, no. 4, pp. 1159–1166, Apr. 2012.

[4] J. N. Laneman, D. N. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3062–3080, Nov. 2004.

[5] M. A. Mulatu, L.-C. Chang, Y. Han Sam, F. G. Mengistu, and D.-F. Tseng, “Threshold-based cooperative communication of energy harvesting active networked tags,” International Journal of Ad Hoc and Ubiquitous Computing, vol. 24, no. 4, pp. 225–236, 2017.

[6] M. A. Mulatu, “Energy cooperation in communication of energy harvesting tags,” AEU-International Journal of Electronics and Communications, vol. 71, pp. 145-151, 2017.

[7] B. Medepally and N. B. Mehta, “Voluntary energy harvesting relays and selection in cooperative wireless networks,” IEEE Trans. Wireless Commun., vol. 9, no. 11, pp. 3543–3553, Nov. 2010.

[8] H. Li, N. Jaggi, and B. Sikdar, “Relay scheduling for cooperative communications in sensor networks with energy harvesting,” IEEE Trans. Wireless Commun., vol. 10, no. 9, pp. 2918–2928, Sep. 2011.

[9] S.-I. Chu, “Performance of amplify-and-forward cooperative diversity networks with generalized selection combining over nakagami-m fading channels,” Communications Letters, IEEE, vol. 16, no. 5, pp. 634–637, 2012.

[10] A. Minasian, S. Shahbaz Panahi, & R. S. Adve, “Energy harvesting cooperative communication systems,” IEEE Transactions on Wireless Communications, vol. 13, no. 11, pp. 6118-6131, 2014.

[11] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, “Signal processing with compressive measurements,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 445–460, Mar. 2010.

[12] M. Puterman, Markov Decision Processes: Discrete stochastic Dynamic Programming. John Wiley and Sons, 2005.