Game Theoretic Anti-jamming Dynamic Frequency Hopping and Rate Adaptation in Wireless Systems

M. K. Hanawal, M. J. Abdel-Rahman, D. Nguyen, and M. Krunz
Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ 85721, USA
{mhanawal, mjabdelrahman, dnnguyen, krunz} @email.arizona.edu

Technical Report
TR-UA-ECE-2013-3

Abstract

Wireless transmissions are inherently broadcast in nature and are vulnerable to jamming attacks. Frequency hopping (FH) and transmission rate adaptation (RA) have been used to mitigate jamming. However, recent works have shown that using either FH or RA (but not both) is inefficient against smart jamming. In this paper, we propose mitigating jamming by jointly optimizing the FH and RA techniques. We consider an average-power constrained "sweep" jammer that aims at degrading the goodput of a wireless link. We model the interaction between the legitimate transmitter and jammer as a *Markov zero-sum game*, and derive optimal defense strategies against worst-case attack strategies. We compare the *average goodput* and *success rate* under the new scheme with the schemes that rely on either FH or RA, but not both. Numerical investigations show that the new scheme improves the average goodput and provides better jamming resiliency.

Index Terms

Dynamic frequency hopping, jamming, Markov decision processes, Markov games, waveform adaptation.

I. Introduction

The convenience of wireless communications and its support for mobility have increased the popularity of wireless devices. While wireless networks offer great flexibility, their open broadcast nature leaves them vulnerable to various security threats, including jamming attacks. In a jamming attack, an adversary injects interfering power into the wireless medium that can hinder legitimate wireless communication in one of two ways: (i) the jamming power can degrade the signal-to-interference-plus-noise ratio (SINR) at a legitimate receiver, and (ii) in carrier sensing networks, continuous jamming may prevent the legitimate transmitter from accessing the medium, hence, creating a denial-of-service attack. In this paper, we consider the attack of the first type and develop optimal defense strategies against it. The open nature of the wireless medium makes it easy for an adversary to monitor communications between wireless devices and use readily available off-the-shelf commercial products to launch a stealth jamming attacks [1], [2], [3].

Several physical-layer techniques have been developed to mitigate jamming, including spread spectrum (e.g., frequency hopping), directional antennas, and power/coding/modulation control. Jammer-specific techniques have also been developed [1]. Common jamming models in the

literature include random jammer, constant jammer, proactive jammer, reactive jammer [2], [4], etc. This classification is based on the channel behavior and transmission capabilities of the jammer. Constant and proactive jammers always emit power into the medium. While a constant jammer transmits power at a fixed level all the time, a proactive jammer can vary the power to meet various constraints. A reactive jammer exhibits more capabilities and emits power only when he detects a legitimate transmission [3]. In this paper, we consider a proactive jamming attack in which the jammer jams one channel at a time and sweeps through all channels. The process is repeated over and over again, but each time with a new (random) sweep pattern. The amount of damage the jammer can inflict depends on his capabilities. Although transmitting continuously at the maximum power will enable the jammer to cause the maximum harm, this happens at the cost of high energy consumption and, more importantly, a high likelihood of being detected. In this work, we assume the jammer has a constraint on its average power.

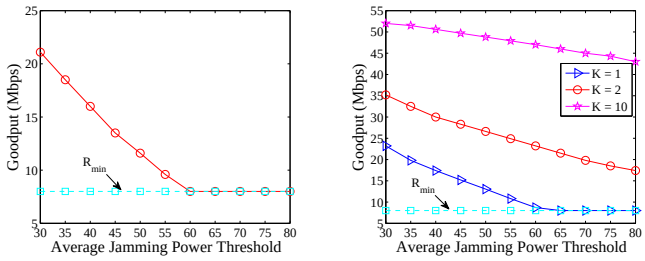
Frequency hopping (FH) [5], [6] and rate adaption (RA) [7], [8] are commonly used techniques to mitigate jamming. However, these techniques are shown to be ineffective when applied separately. In the case of RA with no FH, it is shown that by merely randomizing his power levels the jammer can force the transmitter to *always* operate at the lowest rate [9] if the jamming average power reaches a given threshold. This is captured through a zero-sum game whose Nash Equilibrium's (NE) throughput is plotted in Figure 1(a). Experiments on IEEE 802.11 networks with different RA schemes (e.g., SampleRate, ONOE) [10], [11] also confirm this observation. FH is shown to be largely inadequate in coping with jamming attacks in current 802.11 networks [12]. In particular, when the number of channels is small and channels are not perfectly orthogonal, experimental studies in [12] show that the jammer can degrade the link goodput significantly.

Our aim in this paper is to study the effectiveness of a jointly optimized RA and FH technique to mitigate jamming. The potential of jointly optimizing the RA and FH to combat jammers is demonstrated in Figure 1. In this figure, we draw the NE's throughput of the zero-sum game between a jammer and a link that has freedom in selecting the transmission rate and hopping between channels. In this work, we will develop mechanisms that guide the link to make instantaneous decision (i.e., *pure strategy*)[13] with or without complete information of the jammer.

In a multi-channel system, the transmitter can run away from the jammer by hopping from one channel to another. But, hopping may result in loss of throughput as transmitter may not be able to start transmission on the new channel instantaneously. Also, it cannot reside on the same channel for longer time as the sweep jammer may reach that channel. In adopting transmission rates the transmitter faces similar dilemma; Using higher rate increases chances of getting jammed. On the other hand, if it uses lower rate, it will encounter a loss in throughput. We seek to derive jointly optimal *dynamic frequency hopping* and *rate adaptation* policies for the transmitter against a proactive sweep jammer. This policy informs the transmitter when to switch (hop) to another channel and when to continue (stay) on the current channel. Furthermore, it gives the best rate to use in both cases (hop and stay).

Main Contributions:

- We model the interaction between a legitimate transmitter and a power-constrained sweep jammer as a Markov zero-sum game. The transmitter dynamically decides when to switch the operating channel and what transmission rate to use.
- The optimal defense strategy of the transmitter is derived using Markov decision processes (MDPs), and the structure of the optimal policy is shown to be threshold type. We analyze



(a) $K = 1$ (b) $K > 1$

Figure 1. Single-channel RA vs. multi-channel RA.

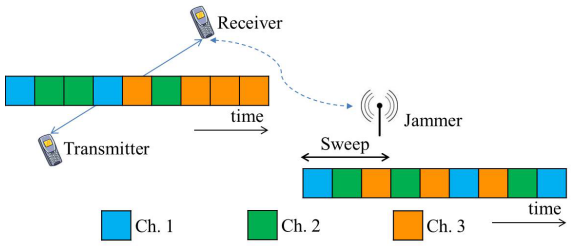


Figure 2. System model.

the "constrained-Nash equilibrium" of the Markov game and show that the equilibrium defense strategy of the transmitter is deterministic.

- We compare the average goodput and the success rate (percentage of un-jammed transmissions) under the proposed jointly optimized RH and RA technique with adaptive FH and RA techniques, considered separately. Through numerical investigations, we show that the new scheme better average goodput and success rate for any set of parameters.

Paper Organization–The rest of the paper is organized as follows. In Section II, we present the channel model, jammer and transmitter models, and the attack and defense models. In Section III, we develop a repeated game to model the interactions between the transmitter and the jammer. In Section IV, we study the Markov zero-sum game and derive optimal defense strategies (hopping and rate adaptation). Numerical results are provided in Section VI. Finally, we conclude the paper in Section VII.

II. System Model

Consider a legitimate transmitter that communicates with its receiver in the presence of a jammer, as shown in Figure 2. The transmitter can communicate on any one of K available channels in each time slot. Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ denote the set of non-overlapping channels. Each channel experiences additive white Gaussian noise (AWGN) with a fixed noise variance. For simplicity, we assume that the noise variance is the same across all channels, and denote it by σ^2.

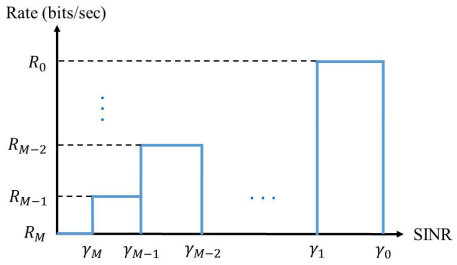


Figure 3. Rate vs. SINR relationship.

A. Transmitter and Channel Model

Time is slotted, and transmissions are assumed to be packet-based, i.e., transmissions happen on disjoint intervals and during these intervals the states of the transmitter and the jammer remain unchanged. The jammer injects additive interference into the channels to degrade the SINR. On each channel, the rate achieved by the transmitter depends on the SINR achieved at the receiver. Consider channel $f_k \in \mathcal{F}$. Let the received power from the transmitter be P_T and that from the jammer be P_J. The SINR at the receiver on channel f_k is given by:

$$\eta_k = \frac{P_T}{P_J + \sigma^2}. \tag{1}$$

For a given SINR, only certain rates are supported at the receiver. The relationship between the achievable rates and the SINR is shown in Figure 3. When the SINR is between γ_i and γ_{i-1}, only rates $R_M, R_{M-1}, \ldots, R_{M-i}$ are achievable. If the transmitter transmits at a rate higher than R_i while the SINR at the receiver is less than γ_i, the transmitted packet is completely lost.

We assume that the transmitter supports $M+1$ different rates (waveforms) on any given channel. The set of rates is denoted by $\mathcal{R} = \{R_0, R_1, \ldots, R_M\}$. Without loss of generality, we assume that $R_0 > R_1 > R_1 > \cdots > R_M$. The transmitter and the jammer are assumed to be equipped with a single radio. The transmitter can communicate on one channel in a given time slot, and can either switch to another channel or stay on the same channel in the next time slot.

B. Jamming Model

We consider a peak-and-average-power-constrained jammer, as in [9] (note that only RA was considered in [9]). The jammer can emit a maximum power of $P_{J,\max}$ in each time slot. The jammer also has a constraint $P_{J,\text{avg}}$ on its average power, where $P_{J,\text{avg}} < P_{J,\max}$. In each time slot, the jammer can choose from $N+1$ discrete power levels $\mathcal{P}_J = \left\{ P_{J_i} = \frac{iP_{J,\max}}{N}, i = 0, 1, \ldots, N \right\}$. We denote $\mathcal{N} = \{0, 1, \cdots, N\}$.

As argued in [9], under an average-power constraint, the attack strategy is to choose a distribution on the set of available powers that satisfies the average power constraint. $\mathcal{P}_J$ denotes the jammer's set of pure strategies. Let J_s denote the strategy space of the jammer and $\mathbb{Y}$ denote an $(N+1)$-probability simplex. Then, $J_s \subset \mathbb{Y}$ and is given by:

$$J_s = \left\{ \mathbf{y} = (y_0, y_1, \ldots, y_N), \sum_{i=0}^{N} y_i = 1, \mathbf{y}\mathcal{P}_J^T \leq P_{J,\text{avg}} \right\}. \tag{2}$$

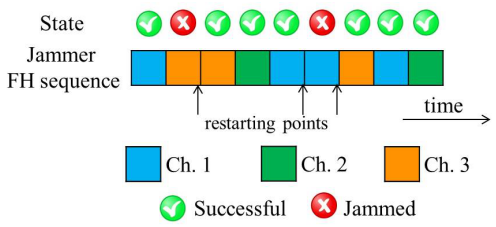


Figure 4. Sweep jammer.

For a given system, the relationship between the rate and the SINR is fixed, as shown in Figure 3. γ_i's in Figure 3 may not correspond to the SINR defined in (1) for a given set of jamming power levels. However, for simplicity, we assume that a one-to-one map exists between the jamming power levels and the highest achievable rates between the transmitter-receiver pair. So when the jammer's power is $P_{J_i}, i = 0, 1, \cdots, N$, the highest achievable rate is R_i. Note that in this mapping, we implicity assume that $N \leq M$. In particular, we will restrict ourselves to the case of $N = M$. Lemma 1 in [9] shows that to analyze the zero-sum game between the transmitter and the jammer, it is enough to consider only the first $\min\{M+1, N+1\}$ entries in $\mathcal{P}_J$ and $\mathcal{R}$.

C. Attack and Defense Strategies

If the jammer (transmitter) knows the strategy of its opponent, then it can come with a better attack (defense) strategy. The strategies depend on the the hardware and computational capabilities of both players. Recall that both transmitter and jammer are each equipped with a single radio, and hence can transmit on one channel per slot. Also, the transmitter and jammer do not have sensing capabilities, i.e., jammer is not aware of the channel used by the transmitter, and vice versa. For simplicity of analysis, we assume that when the jammer successfully jams the transmitter, both of them know that they are on the same channel.

If there is just one channel, i.e., $K = 1$, the only way the transmitter can escape from the jammer is by adapting its rate. In this case, it is shown in [9] that by randomizing its power levels, the jammer can force the transmitter to use the lowest rate. Thus, when multiple channels are available, it is best for the transmitter to hop from one channel to other and evade the jammer. Then, the jammer would also switch to another channel in search of the transmitter.

When the jammer is aware that transmitter may switch its channel, one naive attack strategy for the jammer is to randomly choose one of the K channels with equal probabilities in each time slot. In this case, as argued in [14], the transmitter should stay on the same channel. Anticipating the transmitter's response, the jammer may now go through all the K channels sequentially, jamming one channel in each slot. The jammer could further randomize its sweep pattern each time it successfully jams the transmitter[1] or completes one sweep cycle to make its sweep patterns unpredictable. This type of attack is referred to as *sweep attack* [14] and the jammer is referred to as *sweep jammer*. The sweep jammer is depicted in Figure 4.

[1]otherwise the transmitter can stay on the same channel and transmit successfully in the remaining of the sweep period.

III. DFH GAME WITH RATE ADAPTATION

In this section, we develop a repeated-game model between the sweep jammer and the transmitter, and derive the optimal attack (defense) strategy for the jammer (transmitter). The defense strategy is to decide whether to remain on the same channel or switch to another channel in each time slot, and which rate to use. The attack strategy is to choose a jamming power level in each time slot while satisfying the average-power constraint.

A. Frequency Hopping

The hopping pattern of the sweep jammer is clear– it sweeps all the K channels sequentially, and if the jammer is successful in destroying a transmitted packet, a new random cycle is restarted immediately, without completing the previous cycle, see Figure 4. For the transmitter, hopping is uniform. Each time the transmitter decides to hop, it chooses a channel from the set $\mathcal{F}$ with equal probability. We assume that transmitter does not have a means to know the quality of the channel and assigns no priority to any channel. The transmitter-receiver pair follows a common FH pattern, generated by a pseudo-random sequence. The receiver hops to the same channel as the transmitter if it does not hear from the transmitter for a predetermined time.

B. Reward and Cost

Recall that a transmission at rate R_i is successful if the SINR at the receiver is at least γ_i; otherwise, the packet is completely lost. If the transmitter successfully sends at rate R_i, it obtains a reward of R_i units. If the transmission fails, the transmitter incurs a cost of L units. Specifically, a transmission failure disrupts the communication between the transmitter and the receiver, who need to re-establish their communication through the exchange of several packets, that do not contribute to actual information. Hence, L corresponds to the throughput loss due to a failed transmission. In addition, when the transmitter switches to another channel, it needs to wait till the receiver hops onto the same channel. This waiting period can also result in loss in throughput. We denote the loss in throughput due to channel switching as C units.

In line with [14], we define the transmitter payoff as the difference between the reward and costs it incurs in each time slot. Let $U(n)$ denote the payoff[2] of the transmitter in slot n. The payoff is given as:

$$\begin{aligned} U(n) \quad &= \sum_{i=0}^{M} R_i \cdot \mathbf{1}\big[\text{successful transmission at rate } R_i \text{ in slot } n\big] \\ &\quad - L \cdot \mathbf{1}\big[\text{jamming is successful}\big] - \mathbf{C} \cdot \mathbf{1}\big[\text{transmitter hops}\big], \end{aligned}$$

where $\mathbf{1}[\mathbf{A}]$ is the indicator function. We note that only one term can be positive in the summation term above, as the transmitter can use only one rate in each time slot. An action taken by the transmitter in a given time slot affects its payoff in future time slots. Thus, we will consider a total discounted payoff with a discount factor $\delta \in (0, 1)$, which indicates how much the transmitter values its future payoff over its current payoff. Let $\bar{U}$ denote the total discounted payoff of the transmitter. Then, $\bar{U}$ is given by:

$$\bar{U} = \sum_{n=0}^{\infty} \delta^n U(n). \tag{3}$$

[2]We specify the exact payoff function in the next section after defining state space.

We model the interaction between the transmitter and the jammer as a *Markov* zero-sum game and derive optimal defense and attack strategies for this game. Note that the jammer is constrained on the average power level. We shall derive *constrained Nash equilibria* of the Markov zero-sum game and characterize the properties of the optimal policies using Markov decision processes.

IV. Markov zero-sum game

A Markov game is characterized by an action space, immediate reward for each player, and a state space with the transition probabilities. The decision epochs are at the end of time slot, and the effect takes place in the beginning of the next time slots. The state of the system identifies the status of the transmitter. The state is defined as $x = (x_1, x_2)$, where x_1 denotes the number of contiguous slots since the beginning of the current sweep cycle that the transmitter has been successful and before it last hopped, and x_2 denotes the number of time slots the transmitter is transmitting successfully since it last hopped. Each component takes nonnegative integer values. Note that x_1+x_2 gives the number of time slots the transmitter has been successfully transmitting since the beginning of the current sweep cycle. The state $(0, 0)$ denotes that the transmitter is jammed (zero successful transmissions). Let X denote the state space. Then, X is given by:

$$X = \{(x_1, x_2) : x_1, x_2 = 0, 1, 2, \ldots, K\}. \tag{4}$$

At the end of each time slot, the transmitter has to decide whether to stay on the channel it is currently using, or to hop to a randomly selected channel (which may end up being the same channel). In both cases, the transmitter also has to decide which rate to use from the set $\mathcal{R}$. Therefore, the set of actions available to the transmitter for any state in X is as follows:

$$A = \{(s, R_1), \ldots, (s, R_N), (h, R_1), \ldots, (h, R_N)\} \tag{5}$$

where action (s, R_i) represents the transmitter's decision to stay in the current channel and use rate R_i, and the action (h, R_i) represents the decision to randomly hop to a channel in $\mathcal{F}$ and use rate R_i on that channel. For notational convenience, we write $s_i = (s, R_i)$ and $h_i = (h, R_i)$. Note that we allow the transmitter to stay on a channel, irrespective of whether it got jammed or not on that channel in the previous time slot. We write the transmitter's payoff $U(n)$ as $U_n(x, a, x')$, which denotes the immediate reward for the transmitter when it enters state x' in time slot n after taking action $a \in A$ at the end of time slot $n-1$ while in state x. We assume that this reward is the same in each time slot, and drop the subscripts when there is no ambiguity. For any $(a, x) \in A \times X$, the immediate payoff of the transmitter is defined as follows: $U(\cdot, a, x)$

$$= \begin{cases} -L - C, & \text{if} x = (0,0), a = h_i, \forall i = 1, 2, \ldots, N \\ -L, & \text{if } x = (0,0), a = s_i, \forall i = 1, 2, \ldots, N \\ R_i - C, & \text{if } x \neq (0,0), a = h_i, \forall i = 1, 2, \ldots, N \\ R_i, & \text{if } x \neq (0,0), a = s_i, \forall i = 1, 2, \ldots, N. \end{cases} \tag{6}$$

Note that the reward of transmitter depends only on the action it takes and the new state it enters, and not on its current state. Since the jammer cannot observe the state of the transmitter, it just chooses power levels such that its average power is constrained. For any strategy $\mathbf{y} = (y_0, y_1, \cdots, y_N) \in J_s$ of the jammer, define $Y_i = \sum_{j>i} y_j$ for $i \in \mathcal{N}$.

We next derive the transition probabilities of the Markov chain (see Figure 5). Let $P_{\mathbf{y}}(x'|x, a)$ denote the transition probability to state x' when the current state is x and the transmitter chooses

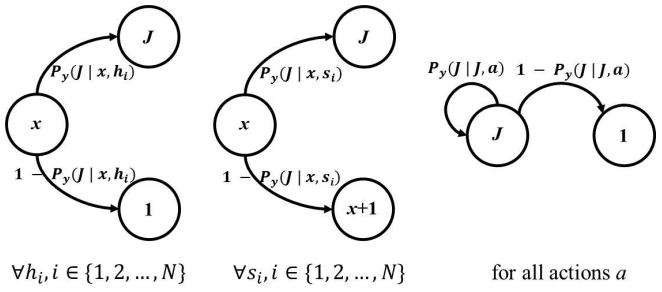


Figure 5. Transition probabilities (J denotes jammed state)

action $a \in A$ while the jammer uses strategy $\mathbf{y}$. When the current state is $(0,0)$, the new state can be either $(0,0)$ or $x = (1,0)$, irrespective of what action is taken by the transmitter. Let $x = (0,0)$. Then, on taking action h_i the system enters into state $(0,0)$ again , only if the jammer also hops onto the same channel as the transmitter and uses a power level that does not allow the transmitter to send at rate R_i. Recall that on each successful jam, the jammer reorders his sweeping pattern independently of his past sweeping pattern. Then, the transmitter and the jammer hop to the same channel with probability $1/K$. On the other hand, if the transmitter decides to stay on the same channel following a successful jamming, the situation will not change either, as the jammer randomly reorders the sweeping pattern. Hence, for $i \in \mathcal{N}$, we have:

$$\begin{aligned} P_{\mathbf{y}}((0,0)/(0,0), h_i) &= Y_i/K = 1 - P_{\mathbf{y}}((1,0)/(0,0), h_i) \\ P_{\mathbf{y}}((0,0)/(0,0), s_i) &= Y_i/K = 1 - P_{\mathbf{y}}((1,0)/(0,0), s_i). \end{aligned} \tag{7}$$

When the current state $x \neq (0,0)$, say $x = (\tilde{x}_1, \tilde{x}_2)$, the next state x' can be either $(0,0)$, $(\tilde{x}_1, \tilde{x}_2 + 1)$, or $(\tilde{x}_1 + \tilde{x}_2, 1)$. Suppose that the transmitter decides to stay on the same channel and use rate R_i. The system then enters into state $(0,0)$ if jammer hops to the same channel and transmits at a power that does not allow packet decoding at rate R_i. Note that the jammer can jam the transmitter on a given channel, say f, provided the jammer has not swept through f in the last $\tilde{x}_1 + \tilde{x}_2$ time slots. In this case, it is also clear that the transmitter knows that f has not been swept by the jammer in the last $\tilde{x}_2$ slots. However, whether or not the jammer swept the channel in the first $\tilde{x}_1$ slots is not known to the transmitter. The transition probability for this case can be computed as the product of the probability that the jammer has not used channel f in the last x_2 slots and the probability that it hops onto f in slot $\tilde{x}_1 + \tilde{x}_2 + 1$ given that it has not swept f in the past $\tilde{x}_2$ slots. If the transmitter decides to hop and use rate R_i, it will go into state $(\tilde{x}_1 + \tilde{x}_2, 1)$ if any of the following happens: (i) the transmitter hops into a channel, say f, that is already swept by the jammer, (ii) f is not swept by the jammer and the jammer does not hop to f in time slot $\tilde{x}_1 + \tilde{x}_2 + 1$, or (iii) the jammer hops onto f and uses a power level that does not disrupt the transmission at rate R_i. We summarize these transition probabilities as follows. For $i \in \mathcal{N}$, $1 \leq x_1 \leq K$, and $1 \leq x_1 + x_2 + 1 \leq K$, we have:

$$\begin{aligned} P_{\mathbf{y}}((0,0)/(x_1, x_2), s_i) &= Y_i/(K - x_2) \\ &= 1 - P_{\mathbf{y}}((x_1, x_2 + 1)/(x_1, x_2), s_i). \end{aligned} \tag{8}$$

$$P_{\mathbf{y}}((x_1+x_2,1)/(x_1,x_2),h_i) =$$
$$1-P_{\mathbf{y}}((0,0)/(x_1,x_2),h_i) = \frac{x_1+x_2}{K}$$
$$+\frac{K-x_1-x_2}{K}\left\{1-\frac{1}{K-x_1-x_2}+\frac{1-Y_i}{K-x_1-x_2}\right\}$$
$$=1-Y_i/K. \tag{9}$$

Remark 1: All transition probabilities depend on the state (x_1, x_2) only through x_2. Henceforth, we denote the state simply as x, which represents the number of successful transmissions on a channel since the transmitter last hopped onto it or since the beginning of the current sweep cycle, whichever occurred later, where $x \in \{0, 1, 2, \ldots, K\}$. With some abuse of notation, we again use X to denote state space, i.e., $X = \{0, 1, 2, \cdots, K\}$.

Let $r_{\mathbf{y}} : X \times A \to \mathbb{R}$ denote the expected reward for the transmitter when the jammer's strategy is $\mathbf{y}$, where $\mathbb{R}$ denotes the set of real numbers. For action a in state x, the expected immediate reward for the transmitter is given by:

$$r_{\mathbf{y}}(x,a) = \sum_{x'} P_{\mathbf{y}}(x'/x,a)U(x,a,x'). \tag{10}$$

In each time slot, the transmitter takes an action that depends on its past observation. For simplicity, we shall restrict ourselves to Markov stationary policies[3], where the transmitter takes action based on his current state only. Let $\pi : X \to A$ denote a decision policy of the transmitter, Π_s denote the collection of stationary policies, and $\pi(x)$ denote the action the transmitter takes in state $x \in X$. For a given $\pi \in \Pi_s$ and a given $\mathbf{y}$, the expected discounted payoff of the transmitter, when the initial state is $x \in X$ is

$$\tilde{V}(x,\pi,\mathbf{y}) = \mathbb{E}^{\pi,\mathbf{y}}\left\{\sum_n \delta^n r_{\mathbf{y}}(X_n,A_n)|\ X_0 = x\right\} \tag{11}$$

where $\{(X_n, A_n) : n = 1, 2, \ldots\}$ is a sequence of random variables, denoting the state-action pair in each time slot. This sequence evolves according to the policy pair $(\pi, \mathbf{y})$. Note that A_n consists of the transmitter's action and the power level chosen by the jammer. The operator $\mathbb{E}^{\pi,\mathbf{y}}$ denotes the expectation over the process induced by the policies π and $\mathbf{y}$.

The transmitter's objective is to choose a policy that results in the highest expected reward starting from any state $x \in X$, defined as

$$V_T(x,\mathbf{y}) = \max_{\pi \in U_s} \tilde{V}(x,\pi,\mathbf{y}). \tag{12}$$

In contrast, the jammer's objective is to choose a strategy $\mathbf{y}$ that minimizes the transmitter's expected discounted payoff, i.e., $\forall x \in X$,

$$V_J(x,\pi) = \min_{\mathbf{y} \in J_s} \tilde{V}(x,\pi,\mathbf{y}). \tag{13}$$

Note that the strategy space of the jammer is constrained, and the transmitter can choose any

[3]For any given history-dependent policy, there exists a Markov policy that is equally good [15][Ch. 4].

stationary policy. We will say a strategy pair $(\pi^*, \mathbf{y}^*)$ is *constrained Nash equilibria* if $\mathbf{y}^* \in J_s$ and

$$\tilde{V}(x, \pi, \mathbf{y}^*) \leq \tilde{V}(x, \pi^*, \mathbf{y}^*) \leq \tilde{V}(x, \pi^*, \mathbf{y}) \tag{14}$$

for all $x \in X$, $\pi \in \Pi_s$, and $\mathbf{y} \in J_s$. Let $V^*(x) \stackrel{\text{def}}{=} \tilde{V}(x, \pi^*, \mathbf{y}^*)$. $\{V^*(x), x \in \mathcal{N}\}$ is referred to as the values of the zero-sum game[4].

Theorem 1: The zero-sum game has a stationary constrained Nash equilibria.

Proof: While the jammer aims to minimize the transmitter's payoff, it needs also to meet its average-power constraint. Since the jammer does not know the value of the current state, its constraint on the average power for any strategy $\mathbf{y}$, i.e., $\mathbf{y}\mathcal{P}^T \leq J_{\text{avg}}$, can be equivalently written as a constraint on an expected discounted cost, as follows:

$$C_\beta(\pi, \mathbf{y}) = (1-\beta)\mathbb{E}^{\pi,\mathbf{y}} \left\{ \sum_{n \geq 0} \beta^n C(X_n, A_n) \right\} \leq J_{\text{avg}} \tag{15}$$

for some $\beta \in (0\ 1)$, where $C(X_n, A_n)$ denotes the cost for the jammer, which is the power it chooses in time slot n. Further, by choosing a strategy $\mathbf{y}'$ such that $y'_0 = 1$, the constraint on the expected discounted cost is strictly met.

Thus, strong Slater condition in [17] is verified and the existence of stationary constrained Nash equilibria follows from Theorem 2.1 in [17]. ■

A. Optimal Defense Strategy at the Transmitter

In this subsection, we derive the properties of the optimal defense strategy for the transmitter against a given jammer's strategy. Let $\pi^*_{\mathbf{y}}(X)$ denote the policy that maximizes the expected discounted reward function when the jammer uses strategy $\mathbf{y}$. For simplicity, we do not explicitly mention this dependency on $\mathbf{y}$, and write $V_T(x, \mathbf{y}) = V(x)$ for notational convenience.

We use the value iteration [15][Ch. 6] method to derive the optimal defence strategy and its properties. The well-known Bellman equations for the expected discounted utility maximization problem in (12) are written as follows:

$$\begin{aligned} Q(x,a) &= r_{\mathbf{y}}(x,a) + \delta \sum_{x' \in X} P_{\mathbf{y}}(x'|x,a) V(x') \\ &= \sum_{x' \in X} P_{\mathbf{y}}(x'|x,a) \left\{ U(x,a,x') + \delta V(x') \right\} \\ V(x) &= \max_{a \in A} Q(x,a). \end{aligned} \tag{16}$$

Note that in our formulation sates $x = 0$ and $x = K$ are equivalent, as the jammer will be starting the sweep cycle afresh at the end of each sweep cycle or after successfully jamming the transmitter. Hence, when the transmitter begins in either state 0 or state K, it should get the same total discounted reward, i.e., $V(0) = V(K)$. From (16), for any $x = 0, 1, \ldots, K-1$, $V(x)$ is expressed in terms of $V(0)$ and $V(x+1)$. Furthermore, because $V(0) = V(K)$, V cannot be a monotone function on the set X. However, we enforce the monotonicity by restricting the transmitters reward in state $K-1$, and use this monotonicity property to establish structure of the optimal policy.

[4]If $(\tilde{\pi}, \tilde{\mathbf{y}})$ is another equilibria, it also results in the same value of the game, i.e., $V^*(x) = \tilde{V}(x, \tilde{\pi}, \tilde{\mathbf{y}})$ [16][Sec. 3.1].

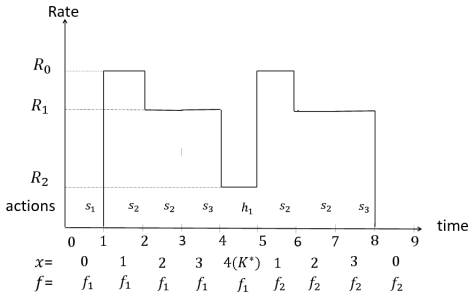


Figure 6. Optimal policy.

Lemma 1: Let $r_{\mathbf{y}}(K-1,a)=0$, for all $a \in A$, i.e., in state $K-1$, the transmitter gets zero reward. Assume $R_N \geq \delta R_0$. Then, $V(\cdot)$ is a decreasing function over $\{0,1,2\cdots,K-1\}$.

From (7) and (9), we note that when the transmitter takes action $h_i, i \in \mathcal{N}$, the probability of entering into state 0 or 1 does not depend on the current state. We make use of this observation and the monotonicity of the function $V(\cdot)$ to derive the following structure of the optimal policy.

Proposition 1: The optimal policy π^* is such that

- There exists constants $K^* \in \{1,2,\cdots,K-1\}$ and $i^* \leq N$ such that

$$\pi^*(x) = h_{i^*} \text{ for } K^* \leq x \leq K-1 \quad \text{and} \quad \pi^*(0) = s_{i^*}.$$

- For any integers x and y, such that $1 \leq x < y < K^*$, and $\pi^*(x) = s_j$ and $\pi(y) = s_k$, satisfy $j \geq k$
- If $r_{\mathbf{y}}(0,s_i)$ is decreasing in index i, then $i^* = 0$.

Proof: The proof idea is as follows. Note that $Q(h_i) = Q(x,h_i)$ does not depend on x for all $i \in \mathcal{N}$. We show that $Q(x,s_i)$ is decreasing in x for all $i \in \mathcal{N}$. Then, for for any $i \in \mathcal{N}$ there exists $x \in X$ such that $Q(x,s_i)$ will be smaller than the largest $Q(h_i)$. Detailed proof is in the appendix. ■

The above proposition tells that when a new sweep cycle begins or the transmitter is jammed on a given channel, the transmitter needs to continue to stay on the same channel until it successfully transmits for K^* consecutive time slots, and hops onto another channel after that. While it stays on a given channel, the transmitter adapts its transmission rate– the transmitter lowers the transmission rate as the number of successive successful transmissions increases. When the transmitter hops, it always uses a fixed rate, which is same as that it uses in the time slot that follows a successful jamming slot. Further, this rate is the maximum rate available (R_0) if $r_{\mathbf{y}}(0,s_i)$ is decreasing in index i.

A typical optimal policy is as shown in Figure 6. The x- axis denotes the time index. The state of the transmitter and the channel used in in each time slot are marked below x-axis. In this policy, on completion of a sweep channel or after being jammed, the transmitter stays on the same channel and uses rate R_0. If successful, it uses rate R_1 in the next time slot. After it succeeds to transmit in slot 3 with rate R_2 and in time slot 4 with rate R_3, it takes the hop decision and transmits with rate R_0. The state changes from 4 to 1 after the hop decision as shown on the x-axis. $K^* = 4$ in this policy.

Note that since the transmitter hops once it reaches state K^*, it never enters into the state that is larger than K^*. Thus, if $K^* < K$, the resulting Markov chain is reducible.

Corollary 1: The threshold K^* is decreasing in L, and increasing in K and C.

Proof: The proof follows by noting that for any $x' > x$, the difference $Q(x', s_i) - Q(x, s_i)$ is an increasing function in L and a decreasing function in K, for all $i \in \mathcal{N}$. $Q(x, h_i)$ is a decreasing function in C for all $i \in \mathcal{N}$ and $x \in X$. This verifies that K^* is an increasing function in C. ■

Next, we return to the study of the Markov game.

B. Constrained Nash Equilibria

In this subsection, we give a method to compute constrained-Nash equilibria of the Markov zero-sum game and study its properties. For a given defense strategy $\mathbf{y} \in J_s$, the following linear program solves the recursive equations in (16) [16][Sec 2.3]:

$$\begin{aligned} \underset{V(x)}{\text{minimize}} \quad & \sum_x V(x) \\ \text{subject to} \quad & V(x) \geq r_{\mathbf{y}}(x,a) + \delta \sum_{x' \in X} P_{\mathbf{y}}(x'|x,a) V(x') \\ & \forall x \in X, a \in A \end{aligned} \tag{17}$$

From Theorem 1, we know that the Markov zero-sum game has a constrained Nash equilibria. We use a non-linear version of the above programming method to compute the equilibria. First, we will develop the necessary notation[5]. Let $\mathcal{M}(A)$ denote the distribution on set A and $f : X \to \mathcal{M}(A)$ denote the strategy of the transmitter. $f(x,a)$ denotes the probability of choosing action $a \in A$ in state $x \in X$. We write $f(s) = (f(s,a), a \in A)$. Similarly, denote the jammer's strategy as $g : X \to \mathcal{M}(P_J)$. Since, jammer does not know the state, $g(x) = \mathbf{y} \in J_s$ for all $x \in X$. Let $r(s,a,p)$ denote the immediate reward for the transmitter when transmitter takes action $a \in A$ and the jammer takes action $p \in \mathcal{P}_{\mathcal{J}}$ in state x, and $P(x'|x,a,p)$ denote the corresponding transition probability of entering into state x'. Write, $R(x) = [r(x,a,p)]_{a \in A, p \in \mathcal{P}_J}$ and $T(x,V) = [\sum_{x'} P(x'|x,a,p)]_{a \in A, p \in \mathcal{P}_J}$ for the reward matrix and the transition probability matrix. $R(x)$ and $T(x,V)$ matrix are defined as in the previous section– $r(x,a,p)$ for each action p is obtained similar to $r_{\mathbf{y}}(x,a)$ without taking expectation with respect to jammer's strategy. Consider the following non-linear program:

$$\begin{aligned} \underset{V_1(x), V_2(x)}{\text{minimize}} \quad & \sum_x V_1(x) + V_2(x) \\ \text{subject to} \quad & V_1(x)\mathbf{1} \geq R(x) g^T(s) + \delta T(x, V_1) \quad \forall x \in X \\ & V_2(x)\mathbf{1} \geq -f(s) R(x) + \delta T(x, V_2) \quad \forall x \in X \\ & \mathcal{P}_J g^T(x) \leq J_{\text{avg}} \quad \forall x \in X \end{aligned} \tag{18}$$

Theorem 2: Let $(V_1^*(x), V_2^*(x), f^*(x), g^*(x))$ denote the global minimum of the non-linear program (18). Then, $(f^*(x), g^*(x))$ denotes the optimal constrained Nash-equilibria of the game.

Proof: The nonlinear program (18) is the same as in [16][Sec. 3.7] with the additional average-power constraint on the jammer's strategy. The proof follows from [16][Th. 3.7.2]. ■

Note that though the optimal strategy of the transmitter for a given strategy of the jammer is deterministic, the equilibrium strategy may not be deterministic [16][Ch. 2]. The strategy $g^*(x)$ is

[5]We follow the notational conventions in [16].

the same for all $x \in X$ as jammer does not know the state. We know from the previous subsection that optimal transmitter's strategy against any given $\mathbf{y}$ is deterministic. Thus at equilibrium, the strategy of the transmitter is deterministic, and can be computed by solving (16) for the strategy $\mathbf{y}^* = g^*(s)$ obtained from (18).

V. Minimax Q-learning

In this section we develop a reinforcement algorithm for the transmitter to learn the optimal defense strategy without explicit knowledge of the jammer. We adapt the Minimax Q-learning algorithm for the Markov zero-sum games in [18].

If the receiver can listen on all the channels simultaneously, then the receive can quickly learn the strategy of the jammer by measuring proportion of the times each power level is used by the jammer and communicate it to the transmitter through a feedback channel. In this case, the transmitter's optimal defense strategy is derived in the previous section. However, if the receiver can listen only one channel at a time, then action taken by the jammer (power level) is known[6] only when both the transmitter and the jammer are on the same channel.

Though the optimal strategy of the transmitter for a given strategy of the jammer is deterministic, the equilibrium strategy may not be deterministic [16][Ch. 2]. Minimax Q-learning is a value-function based reinforcement-learning algorithm specifically designed for the zero-sum games. The value of the game when there is no constraint on the jammers strategy is given by

$$\begin{aligned} Q(x,a,p) &= r(s,a,p) + \delta \sum_{x'X} P(x'|x,a,p)V(x') \\ V(x) &= \max_{\pi(x,\cdot)\in\mathcal{M}(A)} \min_{p\in\mathcal{P}_\mathcal{J}} \sum_{a\in A} f(x,a)Q(x,a,p). \end{aligned} \tag{19}$$

The minimax Q-learning algorithms chooses an action in proportion to the reward accumulated from that choice in the past. Specifically, the *quality* of choosing an action in each state is estimated by replacing (19) with

$$\begin{aligned} Q_n(X_n,a,p) &= (1-\mu_n)Q_{n-1}(X_n,a,p) \\ &\quad +(1-\mu_n)(r(s,a,p)+\delta V(X_{n+1})), \end{aligned}$$

where α_n denotes the learning rate.

The mininimax Q-algortihm is presented in Algorithm 1. If we set for every state-action pair

$$\mu_n := \mu_n(x,a,p) = \frac{1}{(1+\text{number of updated for } Q(x,a,p))},$$

it converges to the optimal strategy of the transmitter from [19][Th. 4]. Clearly, learning rate for any state-action pair satisfies $\sum_n \mu_n = \infty$ and $\sum_n \mu_n^2 < \infty$.

Proposition 2: The minimax Q-learning update rule converges to the optimal Q-function with probability 1, provided each state-action pair is infinitely visited.

In step-2, the algorithm deviates from the optimal policy with probability *explor* and chooses an action uniformly at random. This step ensures that the state-action space is adequately explored. Linear programming can be used to obtain optimal policy $\pi(s,\cdot)$ in Step-4.

[6]Since the transmitter transmits at a fixed power, receiver can measure the amount of interference in received signal (e.g., slow fading).

Note that the unlike the Q-learning algorithm presented in [14], the minimax Q-learning algorithm maintains the table for all the possible actions of the jammer. In our case, the transmitters knows the action of the jammer only when they are on the same channel. When they are not on the same channel, the transmitter updates the $Q(s, a, p)$-table assuming that jammer used action P_0 in that time slot.

The policy learned by the minimax Q algorithm converges to the optimal policy, and this policy is learned in total ignorance of the jammer. Thus, the learned strategy will be optimal against any attack by the jammer, and in particular against the sweep attack.

Step-1:Initialization;
For all $x \in X, a \in A, p \in P_J$;
Let $Q(x, a, p) = 1, V(x) = 1, \pi(x, a) = 1/|A|$, and $\alpha = 1$;
for $n = 1, 2, \cdots$ **do**
  Step-2: Choose an action;
  With probability *explor*, return $a \in A$ uniformly;
  Otherwise if in state X_n, return a with probability $\pi(s, a)$;
  step-3:Learn;
  After receiving reward r for moving from X_n to X_{n+1} via action a and opponent's action p;
  if $a = s_i$, *for some* $i \in \mathcal{N}$ **then**
    $Q_n(X_n, s_i, p) = (1 - \alpha)Q_{n-1}(X_n, s_i, p) + (1 - \alpha)(r + \delta V(X_{n+1}))$;
    $Q_n(X, a, p) = Q_{n-1}(X, a, p)$ for other (X, a, p);
  else
    $\forall\ x \in X\ , Q_n(x, a, p) = (1 - \alpha)Q_{n-1}(x, a, p) + (1 - \alpha)(r + \delta V(X_{n+1}))$;
    $Q_n(X, a, p) = Q_{n-1}(X, a, p)$ for other (X, a, p)
  end
  Step-4:Update Policy;
  if $|V - V'| \leq \epsilon$ **stop**; **else**
  choose $\pi(x, \cdot)$ for all $x \in X$ such that;
  $\pi(x, \cdot) = \arg \max_{\pi'(x,\cdot)} \min_{p' \in \mathcal{P}_J} \sum_{a' \in A} \pi'(x, a')Q_n(x, a', p')$;
  $V'(x) = \min_{p' \in \mathcal{P}_J} \sum_{a' \in A} \pi(x, a')Q_n(x, a', p')$;
  $\alpha = \mu_n, V = V'$
end

Algorithm 1: Minimax Q-learning Algorithm

VI. Performance Evaluation

In this section, we study the performance of the proposed zero-sum Markov game under different values of the system parameters. Our performance metrics are the average goodput (in Mbps), the success rate (defined as percentage of un-jammed transmissions), and the hop rate (defined as the rate at which the transmitter switches channel). All performance measures are computed over 1000 sweep cycles. The parameters of study include the number of channels K, the jamming cost L, and the switching cost C. For the jammer we set $J_{\max} = 25$ and $J_{\text{avg}} = 20$. We followed the rate adaptation system of the IEEE 802.11a protocol [20], which uses rates 6, 9, 12, 18, 24, 36, 48, and 54 Mbps. The proposed game is implemented in MATLAB. The 95% confidence intervals are shown in the numerical figures. When they are very tight, they are

not drawn to prevent cluttering the graph. We obtain the joint-optimal defense strategy of the transmitter and attack strategy of the jammer by solving (18).

The proposed joint-optimal policy is compared with three other policies: (i) Optimal FH: transmitter switches channels according to the joint-optimal policy, but transmits at fixed rate of 54 Mbps, (ii) Optimal RA: transmitter adapts rate according to the joint-optimal policy, but does not switch channels and (iii) Random FH: transmitter hops in each slot, and transmit at fixed rate of 54 Mbps. In the random FH strategy, in each slot the transmitter uniformly selects one channel, and does not perform any RA.

A. Average Goodput

Figure 7 depicts the average goodput vs. L for two values of C. Figure 7(a) shows that when C is large (in our example, $C = 30$ Mbps), there is a threshold L^* on the value of L; if $L < L^*$ the optimal FH scheme works almost as good as the joint-optimal scheme, and if $L > L^*$ the performance of the optimal FH scheme starts degrading with L. This shows the importance of RA when L is large, given that C is also large. Figure 7(a) also shows that, in contrast to the optimal FH scheme, the optimal RA scheme is worse than the joint-optimal when L is smaller than a certain threshold (not necessarily the same threshold as L^*), and it behaves almost the same as joint-optimal when L is sufficiently large. The random FH scheme has the worst performance. This is because of the incurred switching cost due to continuous hopping. When C is relatively small (e.g., $C = 5$ Mbps), Figure 7(b) shows that the optimal FH behaves almost the same as the joint-optimal, irrespective of the value of L.

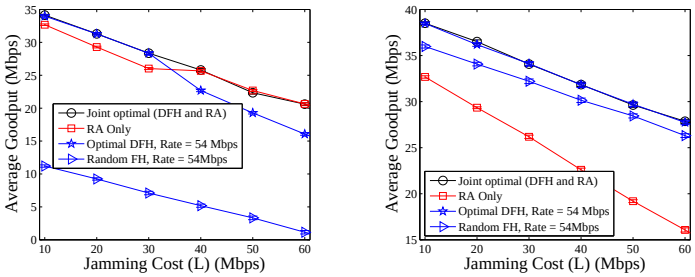
The average goodput is plotted in Figure 8 vs. K. When K is sufficiently large ($K > 4$ in our setup), the optimal FH scheme works almost as good as the joint-optimal scheme, however, if $K < 4$ the joint-optimal works much better than the optimal FH. In contrast, the optimal RA scheme works close to the joint-optimal scheme when K is small, and its performance degrades compared to the joint-optimal scheme when K is large. As shown in Figure 8, the average goodput improves with K.

B. Success Rate

Figure 9 plots the success rate vs. L for two values of C ($C = 30$ Mbps and $C = 5$ Mbps). When C is small, The success rates of both the optimal FH and the joint-optimal schemes are almost the same, and they are non-decreasing with L. The success rate of the random FH scheme is $\sim 100\%$.

When C is large, the success rate of the optimal FH scheme decreases with L, and the achieved improvement in success rate by jointly optimizing the becomes obvious when L is large (joint-optimal has a success rate of $\sim 89\%$ compared to a success rate of $\sim 67\%$ for the optimal FH when $L = 100$ Mbps). When L is sufficiently large, success rate of optimal RA is close to success rate to the joint-optimal scheme.

When $K = 1$, the success rate of both the joint-optimal and the optimal RA scheme is 100%. This is because the transmitter and jammer are both on the same channel and the RA strategy will enforce the transmitter to use the lowest rate (i.e., 6 Mbps. See Figure 7) in order to avoid jamming. The success rate of both schemes drops significantly when K is 2 because in this case the jammer starts sweeping between two channels and the transmitter does not know which channel is currently used by the jammer. When $K > 2$, the success rates of both schemes start increasing with K, but the success rate of the joint-optimal increases faster. For the optimal FH



(a) $C = 30$ Mbps (b) $C = 5$ Mbps

Figure 7. Average goodput vs. L for different values of C ($K = 5$).

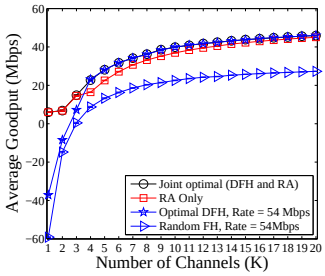


Figure 8. Average goodput vs. K ($L = 40$, $C = 22$).

strategy, the success rate is zero when $K = 1$ because the transmitter and jammer are both on the same channel, and the success rate increases when K increases.

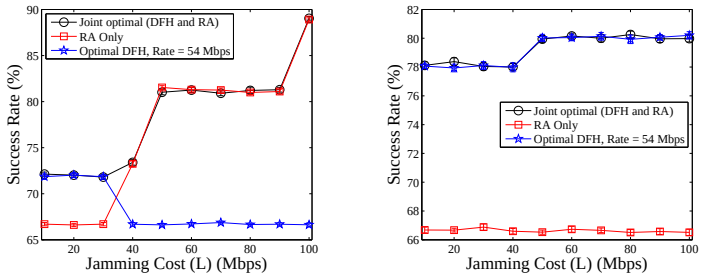
C. Hop Rate

Figure 11 depicts the hop rate vs. L for two values of C. When C is large, both joint-optimal and the optimal FH prefer not to hop when L exceeds a certain threshold, and hop with a low rate ($< 7\%$) when L is small enough. When C is small, both schemes prefer hopping frequently, and the hopping rate increases when L exceeds a certain threshold. Figure 12 plots the hop rate vs. K. Both schemes prefer not to hop for small values of K. When K exceeds a certain threshold, the hop rate of both schemes increase significantly, beyond which the hop rate decreases with K because when K increases the likelihood of meeting the jammer on the same channel decrease, therefore the hop rate decreases.

VII. Conclusions

In this paper, we analyzed a defense strategy against a sweep jammer that is derived by jointly optimizing frequency hopping and rate adaptation techniques. We modeled the interaction between the transmitter and jammer as a Markov zero-sum game and derived optimal equilibrium defense strategy against the worst attack strategy.

Performance evaluation of the proposed scheme shows that the joint-optimal strategy improves improves performance. The joint-optimal scheme is particularly efficient when the number of channels is low. When the number of channels is high, optimal FH performs close to joint optimal



(a) $C = 30$ Mbps (b) $C = 5$ Mbps

Figure 9. Success rate vs. L for different values of C ($K = 5$).

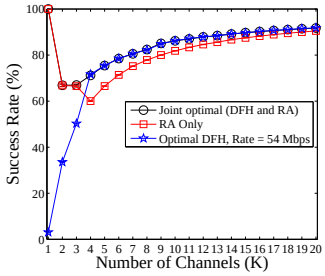
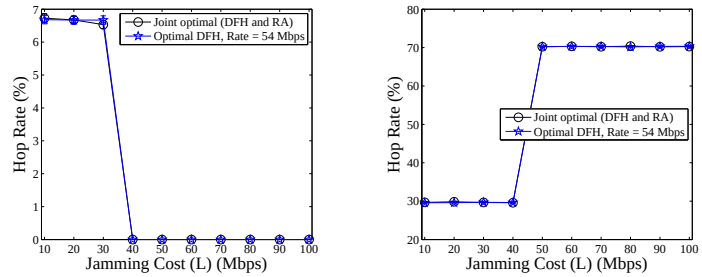


Figure 10. Success rate vs. K ($L = 40$, $C = 22$).



(a) $C = 30$ Mbps (b) $C = 5$ Mbps

Figure 11. Hop rate vs. L for different values of C ($K = 5$).

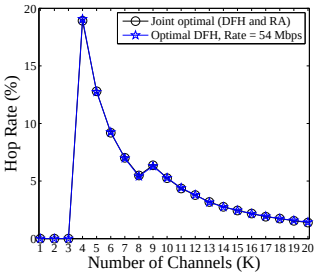


Figure 12. Hop rate vs. K ($L = 40$, $C = 22$).

policy. Numerical evaluations show that threshold effect exists on the performance of the optimal FH and optimal RA schemes. For example, when L and C are high, optimal RA performs better than optimal FH, but as the value of L decreases optimal FH gives better performance than optimal RA. Whereas, the joint optimal scheme gives performance (goodput, success rate, and hop rate) better than both optimal FH and optimal RA for any values of L, C and K. Thus, defense policy derived by jointly optimizing FH and RA techniques achieve better performance over all the system parameters.

REFERENCES

[1] W. Xu, W. Trappe, Y. Zhang, and T. Wood, "The feasibility of launching and detecting jamming attacks in wireless networks," in *Proc. of the ACM MobiHoc Conf.*, Urbana-Champaign, IL, USA, 2005.

[2] S. Khattab, D. Mosseé, and R. Melhem, "Jamming mitigation in multi-radio wireless networks: Reactive or proactive?" in *Proc. of the ACM SecureComm Conf.*, Istanbul, Turkey, September 2008.

[3] R. Gummadi, D. Wetherall, B. Greenstein, and S. Seshan, "Understanding and mitigating the impact of RF interference on 802.11 networks," in *Proc. of the ACM SIGCOMM Conf.*, Kyoto, Japan,, 2007.

[4] E. Bayraktaroglu, C. King, X. Liu, G. Noubir, and R. Rajaraman, "On the performcne of IEEE 802.11 under jamming," in *Proc. of the IEEE INFOCOM Conf.*, Phoenix, AZ, USA, April 2008.

[5] W. Xu, T. Wood, W. Trappe, and Y. Zhang, "Channel surfing and spatial retreats: defenses against wireless denial of service," in *Proc. of the ACM WiSe Workshop*, Philadelphia, PA, USA, Ocotber 2004.

[6] V. Navda, A. Bohra, S. Ganguly, and D. Rubenstein, "Using channel hopping to increase 802.11 resilience to jamming attacks," in *Proc. of the IEEE INFOCOM Conf.*, Anchorage, Alaska, USA, 2007, pp. 2526–2530.

[7] K. Pelechrinis, I. Broustis, S. Krishnamurthy, and C. Gkantsidis, "Ares: An anti-jamming REinforcement System for 802.11 networks," in *Proc. of CoNEXT*, Rome, Italy, 2009.

[8] J. Zhang, K. Tan, J. Zhao, H. Wu, and Y. Zhang, "A practical SNR-guided rate adaptation," in *Proc. of the IEEE INFOCOM Conf.*, Phoenix, AZ, USA, April 2008.

[9] K. Firouzbakht, G. Noubir, and M. Salehi, "On the capacity of rate-adaptive packetized wireless communication links under jamming," in *Proc. of the ACM WiSec Conf.*, Tucson, AZ, USA, 2012.

[10] K. Pelechrinis, I. Broustis, S. V. Krishnamurthy, and C. Gkantsidi, "A measurement-driven anti-jamming system for 802.11 networks," *IEEE/ACM Transactions on Networking*, vol. 19, no. 4, pp. 1208–1222, August 2011.

[11] G. Noubir, R. Rajaraman, B. Sheng, and B. Thapa, "On the robustness of IEEE802.11 rate adaptation algorthms againt smart jamming," in *Proc. of the ACM WiSec conf.*, Hamburg, Germany, June 2011.

[12] K. Pelechrinis, C. Koufogiannakis, and S. Krishnamurthy, "Gaming the jammer: Is frequency hopping effective?" in *Proc. of the ACM WiOpt Conf.*, Seoul, Korea, June 2009.

[13] H. Gintis, *Game Theory Evolving (Second Edition)*. Princeton University Press, 1990.

[14] Y. Wu, B. Wang, K. J. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 1, pp. 4–15, January 2012.

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

[16] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York, USA: springer-Verlag, 1997.

[17] E. Altman and A. Shwartz, "Constrined Markov games: Nash equilibria," *Advances in Dynamic Games and Applications*, vol. 5, pp. 213–221, 2000.

[18] M.L.Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. of the Int'l Conf. on Machine Learning (ICML)*, July 1994, pp. 157–163.

[19] ——, "Value-function reinforcement learning in Markov games," *Journal of Cognitive Systems Research*, vol. 2, pp. 55–66, 2001.

[20] "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std. 802.11*, June 2007.

VIII. APPENDIX

A. Proof of Lemma 1

When in state $K-1$, the transmitter gets zero reward and enters into state K. From 16 we have $V(K-1) = \delta V(K) = \delta V(0)$, and for $x = 0, 1, 2, \cdots K-2$ and $i = 1, 2, \cdots, N$,

$$V(x) \geq \frac{Y_i}{K-x}(-L + \delta V(0)) + \left(1 - \frac{Y_i}{K-x}\right)(R_i + \delta V(x+1)).$$

First note that by taking $i = N$, $V(x) \geq R_N + \delta V(x+1)$. Hence, if we start with positive values for V in the value iteration algorithm the new V is also positive. Implying that $V(x) \geq 0$ for $x \in X$. Consider the difference

$$\begin{aligned} V(K-2) - V(K-1) \geq \frac{Y_i}{2}(-L + \delta V(0)) \\ + \quad \left(1 - \frac{Y_i}{2}\right)(R_i + \delta V(k-1)) - V(K-1). \end{aligned}$$

For $i = N$, we have

$$\begin{aligned} & V(K-2) - V(K-1) \\ & \geq \quad R_N + \delta V(K-1) - V(K-1) \\ & = \quad R_N - \delta(1-\delta)V(0) > 0. \end{aligned} \tag{20}$$

We use the relation $V(K-1) = \delta V(0)$ in the last step. The maximum reward the transmitter can get is R_0. Then, from (12), we get $V(x) \leq \frac{R_1}{1-\delta}$ for $x = 0, 1, 2, \cdots K-1$. The final positivity claim follows by using the condition $R_N \geq \delta R_0$ Following similar steps we can establish the monotonicity in other components.

B. Proof of Proposition 1

As the transition probabilities and the reward do not depend on the state when the transmitter decided to hop, for all $i \in \mathcal{N}$ and $x \in X$, we have

$$Q(x, h_i) = Q(0, h_i) = Q(0, s_i) - CY_i/K \leq Q(0, s_i). \tag{21}$$

It is clear that in state $x = 0$, 'stay' is the optimal strategy. Let $Q^* \stackrel{\text{def}}{=} \max_{i \in \mathcal{N}} Q(0, h_i)$. Then, $Q^* < Q(0, s_i)$. Note that $r_{\mathbf{y}}(0, s_i) = -L(Y_i/K) + R_i(1 - Y_i/K)$ need not be monotone in index i, as both R_i and Y_i are decreasing in i. By applying (16) for all $i \in \mathcal{N}$, we have

$$V(0) \geq r(0, s_i) + \frac{Y_i}{K}\delta V(0) + \left(1 - \frac{Y_i}{K}\right)\delta V(1) \tag{22}$$

with equality for some $i \in \mathcal{N}$. Let m denote this index. Rewriting the above relation with equality for action s_m, we have

$$\begin{aligned} & \frac{Y_m}{K}(R_m + L) = \\ & \quad (R_m - (1-\delta)V(0)) + \delta\left(1 - \frac{Y_m}{K}\right)(V(0) - V(1)). \end{aligned} \tag{23}$$

The left-hand quantity is positive, and the last quantity on the right side is positive by the fact that $V(0) \geq V(1)$. This implies that it must be the case that $R_m \geq (1-\delta)V(0)$. We use this relation to show that $Q(x, s_m)$ is monotonically decreasing in x on the set $\{0, 1, 2, \cdots K-1\}$.

Taking the difference of $Q(x, s_m)$ and $Q(x', s_m)$ for $x \geq x'$, we have

$$\begin{aligned} Q(x, s_m) - Q(x', s_m) = & \\ & \left(\frac{1}{K-x'} - \frac{1}{K-x}\right) Y_m(L + R_m - \delta V(0)) \\ + \delta & \left(1 - \frac{Y_m}{K-x}\right) V(x+1) - \delta \left(1 - \frac{Y_m}{K-x'}\right) V(x'+1) \\ \geq & \left(\frac{1}{K-x'} - \frac{1}{K-x}\right) Y_m(L + R_m - \delta V(0)) \\ + & \left(\frac{1}{K-x'} - \frac{1}{K-x}\right) \delta Y_m V(x'+1) \geq 0. \end{aligned} \tag{24}$$

We used the relation $V(x+1) \geq V(x'+1)$ in the last inequality. Furthermore,

$$Q(x, s_N) - Q(x', s_N) = \delta(V(x+1) - V(x'+1)) > 0. \tag{25}$$

Using similar arguments we can show that $Q(x, s_i)$ is decreasing in x for all $i \in \mathcal{N}$. Indeed, (24) holds for all $i \in \mathcal{N}$ when $R_N \geq \delta R_0$. Since $Q(x, s_i)$ is decreasing in x and $Q^* < Q(0, s_i)$, for all $i \in \mathcal{N}$, there exists an integer K^*, $0 < K^* \leq K-1$, such that $Q(K^*-1, s_i) < Q^* \leq Q(K^*, s_i)$ for all $i \in \mathcal{N}$. In the extreme case where $Q(K-1, s_i) \geq Q^*$ for all $i \in \mathcal{N}$, the transmitter always stays on the same channel.

The second part of the proposition follows by noting that the difference $Q(x, s_i) - Q(x', s_i)$ for any $x > x'$ is increasing in index i. To prove the last part of the proposition, first note that Y_i is decreasing in index i. Due to the fact that $V(0) \geq V(1)$, the sum

$$r_{\mathbf{y}}(0, s_i) + \frac{Y_i}{K} V(0) + \left(1 - \frac{Y_i}{K}\right) V(1) \tag{26}$$

is also decreasing in index i. Hence $Q^* = Q(x, h_1)$ and $V(0) = Q(0, s_1)$. This completes the proof.