Cognitive Radar Applied To Target Tracking Using Markov Decision Processes

Ersin S. Selvi

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical Engineering

R. Michael Buehrer, Chair

Alan J. Michaels, Co-Chair

Allen B. MacKenzie

December 15, 2017

Blacksburg, Virginia

Keywords: Cognitive radar, target tracking, Markov decision process, interference mitigation, spectrum coexistence

Cognitive Radar Applied To Target Tracking Using Markov Decision Processes

Ersin S. Selvi

Academic Abstract

The radio-frequency spectrum is a precious resource, with many applications and users, especially with the recent spectrum auction in the United States. Future platforms and devices, such as radars and radios, need to be adaptive to their spectral environment in order to continue serving the needs of their users. This thesis considers an environment with one tracking radar, a single target, and a communications system. The radar-communications coexistence problem is modeled as a Markov decision process (MDP), and reinforcement learning is applied to drive the radar to optimal behavior.

Cognitive Radar Applied To Target Tracking Using Markov Decision Processes

Ersin S. Selvi

General Audience Abstract

The radio-frequency electromagnetic spectrum is a precious resource, in which users and operators are assigned frequency slots in which they can operate. The federal spectrum auction in the United States freed up some of the spectrum for shared use. The implications of this are the spectrum will become more dense; there will be more devices and users in the same amount of spectrum. The devices and platforms of this spectrum need to be more adaptive and agile in order to (1) not be interfered by other systems, (2) cause interference to other systems, and (3) continue to meet the needs of users (*e.g.* cell phone users) and operators (*e.g.* military radar). The work presented in this thesis applies Markov decision process and reinforcement learning to solve the problem.

Acknowledgments

First and foremost, I would like to thank U.S. Army Research Laboratory, who sponsored the research in this thesis. Specifically from ARL, I would also like to thank Anthony Martone, Kelly Sherbondy, and Amir Zaghloul for their input into the direction of the project, advice and feedback on papers. This project would not be here without their support.

I would like to thank my family for their support while I was in college and graduate school. My parents always pushed me in school, and taught me the value of an education, which I appreciate much more now than ever. To my advisor from Wirless@VT, Dr. R. Michael Buehrer, and my advisor from the Hume Center, Dr. Alan Michaels, both who provided guidance and were always open to answering questions I had about graduate school, electrical engineering, and working in industry; and to Dr. Allen B. MacKenzie, whose Stochastic and Random Processes class proved useful in understanding Markov Decision Processes.

Thanks as well to everyone from the Hume Center, and Wireless@VT labs; especially Dr. William "Chris" Headley, Kayla Brosie, Mahi Abdelbar, Chris O'Lone, Jeff Poston, Tad

Czauski, and Mark Kozy for their advice, guidance, and friendship. Additional thanks to Mario Ortiz, and Amber, Danielle, Dan, and Nikki Thorsen for their warmth and friendship, and for making me feel I had a family while away from home.

Contents

1 Introduction **1**

1.1 Executive Summary 1

1.2 Thesis Overview 4

2 Introduction to Radar **6**

2.1 Radar 6

2.2 Physics of Radar 7

2.3 Antennas 9

2.4 Waveforms 10

2.5 Measured Parameters 11

2.6 Radar Functions 17

2.6.1 Search/Target Detection 17

2.6.2 Target Tracking 18

2.6.3 Imaging 18

2.7 Radar Applications 19

3 Introduction to Cognitive Radar and Machine Learning **22**

3.1 Cognitive Radar Concept and Inspiration 22

3.2 Prior Work in Cognitive Radar 24

3.2.1 Beamforming 24

3.2.2 Target Classification 25

3.2.3 Waveform Optimization and Waveform Diversity 31

3.2.4 Target Tracking 33
3.2.5 Spectrum Sensing and Spectrum Agility 36
3.3 Artificial Intelligence and Machine Learning 37
3.3.1 Reinforcement Learning 38
3.3.2 Markov Decision Processes (MDPs) 39
3.3.3 Summary 43

4 System Model and Detailed Approach **44**
4.1 Proposed System Model 44
4.2 The Radar Environment 45
4.3 Experiment Details 49

5 Experimental Results and Analysis **52**
5.1 Constant interference 53
5.2 Intermittent interference 58
5.3 Triangular frequency sweep 64
5.4 Sawtooth frequency sweep 70
5.5 Length-5 pseudorandom frequency hop 73
5.6 Length-10 pseudorandom frequency hop 75
5.7 Pseudorandom frequency hop 80
5.8 Position-dependent interferer 82
5.9 Comparison With Other Techniques 88

6 Conclusion **92**

Appendix A Analysis of Interference Cases **97**

Appendix B Training and Testing Algorithm **101**

Bibliography **103**

List of Figures

4.1 An example radar scene and trajectory. 46

4.2 Example of the random trajectories used for training. 51

5.1 Results for constant interferer. 56

5.2 Results for 10% intermittent interferer. 61

5.3 Results for 90% intermittent interferer. 62

5.4 Waterfall plot of triangular frequency sweep interferer. 64

5.5 Results for triangle sweep interferer, without memory. 68

5.6 Results for triangle sweep interferer, with memory. 69

5.7 Waterfall plot of sawtooth frequency sweep interferer. 70

5.8 Results for sawtooth frequency sweep interferer. 72

5.9 Results for length-5 frequency hopping interferer. 74

5.10 Results for length-10 frequency hopper, without memory. 78

5.11 Results for length-10 frequency hopper, with memory. 79

5.12 Results for pseudorandom frequency hop interferer. 81

5.13 Trajectory of target with direction-dependent interferer. 84

5.14 Results for direction-dependent constant interferer. 85

5.15 Results for direction-dependent intermittent interferer, with 90% transmission probability. 86

5.16 Results for direction-dependent intermittent interferer, with 10% transmission probability. 87

5.17 Results for comparing MDP and DSA for high intermittent case. 90

5.18 Results for comparing MDP and DSA for triangle frequency sweep case. . . . 91

List of Tables

5.1 Summary of reward structure. 54

5.2 Rewards for each action, when the target is further away from the radar. . . 57

5.3 Rewards for each action, when the target is closer to the radar. 58

5.4 Interference states and actions for triangle sweep interferer, without memory. 66

5.5 Interference states and actions for triangle sweep interferer, with memory. . . 67

5.6 Interference states and actions for sawtooth sweep interferer. 71

5.7 Interference states and actions for length-5 frequency hopping interferer. . . 73

5.8 Interference states and actions for length-10 frequency hopping interferer, without memory. 77

5.9 Interference states and actions for length-10 frequency hopping interferer, with memory. 77

A.1 Value functions for high probability of transmission interference 98

A.2 Value functions for low probability of transmission interference 98

A.3 Value functions for triangular sweep interferer, without memory 99

A.4 Value functions for triangular sweep interferer, with memory 100

Chapter 1

Introduction

1.1 Executive Summary

The radio-frequency electromagnetic spectrum is a precious resource where an abundance of users are competing over finite resources [1]. This spectrum has found uses in radar, communications, radio and television broadcasting, navigation, and sensing [1]. The recent spectrum auction and reallocation [2] has further motivated the need for more effective spectrum sharing technologies [1], between systems and devices of the same application, or even different applications such as radars and communications systems. The concept of a "more intelligent" communication system was introduced by Mitola nearly 20 years ago, in which the *cognitive radio* was envisioned to be able to manipulate its parameters and settings to best serve the needs of its users while also coexisting with other communications systems [3].

In a similar respect, *cognitive radar* has emerged as a potentially powerful solution to solve the challenges facing radar today [4]. Traditional/contemporary radars are designed based on predetermined targets for signal-to-interference-plus-noise ratio (SINR) and maximum operating range, "with target and clutter models that represent averaged, anticipated responses [5]." The resulting design uses fixed (or sets of fixed) parameters, and lacks flexibility in adapting to varying target and environment conditions [5]. When there are variations in the target or environment that depart from the assumed design conditions, the radar's performance will be suboptimal [5]. The traditional radar can only achieve optimal performance in one scenario (the scenario for which it was designed), but is unable to achieve optimal performance over all possible scenarios. Cognitive radar aims to free traditional radars from these restrictions, allowing them to perform optimally across *all* scenarios.

Contemporary research into cognitive radar is generally split into two thrusts: 1. Enhanced radar functionality and performance, and 2. Spectrum sharing. Of interest to this work is work in spectrum sharing; prior works in this thrust include developing policies for coexistence, coexistence between rotating radars and nearby cellular communications systems, and modifying center frequency and bandwidth to avoid interference.

This work proposes modeling the target tracking and radar-communications coexistence problem using a modification of the perception-action cycle and cognitive radar framework discussed in [6]. The perception-action cycle is one of the components of Fuster's paradigm of (biological) cognition [7,8]; *sensors* and *processors* are used to develop a *perception* of the environment, which is then used to take an *action*. The *action* will have some measurable effect on the environment, which will again be *sensed* and *processed* to form a new *perception*, on which a new *action* will be taken [6]. This process repeats as a cycle; the "sensory or internal signals lead to actions that generate feedback that regulates further actions, and so on [8]." The perception-action cycle works reciprocally with *memory* [6]; the *memory* stores the experiences from which the radar can learn from and make new decisions.

The model presented in this work uses the Markov Decision Process and reinforcement learning to learn actions which mitigate interference between the radar and communication systems while optimizing radar performance. Markov Decision Processes (MDPs) model sequential decision problems in which "an agent's utility depends on a sequence of decisions [9]." The goal of this application is to enable the radar to learn from offline training data instead of having to perform online optimization during each radar cycle. The motivation for using MDPs is based on the fact that most communications systems can be modeled as finite state machines [10]. Further, the reward structure of MDPs is flexible, allowing system designers to emphasize interference avoidance or tracking performance as desired. The perception-action cycle manifests as the instantaneous rewards, which evaluates actions taken by the radar and its effect on the environment. The memory manifests as the reward and transition probability functions, which summarizes all of the data the radar has seen during training.

1.2 Thesis Overview

Chapter 2 discusses the fundamentals of radar, namely the physics behind radar operation, antennas, useful information that can be gathered via radar, the main functions of radar, and a list of various applications.

Chapter 3 introduces cognitive radar, and discusses artificial intelligence and machine learning. Specifically, we discuss reinforcement learning, the subfield of artificial intelligence relevant to this work. Markov Decision Processes (MDPs), which are used to model the radar-communications coexistence problem are also presented and discussed.

Chapter 4 presents the system model and explains the setup of the radar environment. The system model involves a single tracking radar, one communications system, and a target. The radar is attempting to maintain the target track, while also avoiding interference caused by the communications system. We also discuss in detail the experimental set up used in this work.

Chapter 5 discusses the experiments in more detail, and presents the results. The results represent several models of interference, including: (1) Constant interference, (2) Intermittent interference, (3) Triangular frequency sweep, (4) Sawtooth frequency sweep, (5) Pseudorandom frequency hopping, and (6) Direction-dependent interference. The results broadly demonstrate that using an MDP-based model and reinforcement learning, the radar can learn the interference behavior, anticipate its spectral occupancy, and adapt its waveform to optimize performance.

This work resulted in a conference paper submitted to the 2018 IEEE Radar Conference, and a journal paper submitted to IEEE Transactions on Aerospace and Electronic System's special session on dynamic spectrum systems.

Chapter 2

Introduction to Radar

2.1 Radar

Radar[1] is an instrument that uses the transmission and reception of radio waves to determine information about a target of interest. Radars transmit electromagnetic (EM) radio-frequency (RF) waves which reflect off the target, and the reflected waves are then received and processed by the same radar system. Any radar system has the following elements: (1) a transmitter, (2) at least one antenna, and (3) a receiver.

Monostatic radars have the transmitter and receiver collocated and sharing the same antenna. Bistatic radars have the transmitter and receiver located a considerable distance from each other, and using different antennas. Multiple-Input, Multiply-Output radar systems are composed of two or more monostatic or bistatic systems working in conjunction. There are various tradeoffs between the setups. By virtue of their setup, monostatic systems

[1] For more about radar, the reader is referred to the following sources: [11–13]

will have fewer components and will thus cost less. However, bi-static and MIMO configurations afford greater capability such as better detection of stealthy targets, but come at a greater cost.

2.2 Physics of Radar

An electromagnetic (EM) wave transmitted by a radar is a coupled pair of oscillating electric and magnetic fields. The electric and magnetic fields are perpendicular to each other, and the plane wave created by the fields is perpendicular to the direction of propagation. The shape traced out by the electric field component describes its polarization. There are several kinds of polarization: horizontal, vertical, circular, elliptical, and random/none. The selection of polarization type will depend on the application.

Although the EM wave has coupled electric and magnetic fields, only the electric field component is utilized for analysis. Electric fields are described by the equation

$$E = E_0 \cos(kz - \omega t + \phi) \tag{2.1}$$

where E_0 is the electric field amplitude, k is the wavenumber, z is the vector in the direction of propagation, ω is the angular frequency, t is time, and ϕ is the phase offset. The wavenumber is equal to $2\pi/\lambda$, where λ is the wavelength. Angular frequency is equal to $2\pi f$, where f is the frequency. The wavelength and frequency of a wave are related by

$$v_p = \lambda f \tag{2.2}$$

where v_p is the phase velocity of the wave. Phase velocity depends on the properties of the propagation medium, and is typically less than or equal to the speed of light in a vacuum, $c \approx 3 \times 10^8$m/s.

The interaction of EM waves with the surrounding environment varies with frequency [11]. For example, the Friis transmission equation - which describes the received power in a communications system link - is defined as

$$P_R = P_T G_T G_R \left(\frac{\lambda}{4\pi R} \right)^2 \tag{2.3}$$

where

P_R = Received power (W)

P_T = Transmitted power (W)

G_T = Transmitter gain (unitless)

G_R = Receiver gain (unitless)

λ = Wavelength (m)

R = Range from transmitter to receiver (m).

Within this equation is the free-space propagation loss, L_P, which is equal to

$$L_P = \left(\frac{4\pi R}{\lambda} \right)^2 = \left(\frac{4\pi R f}{c} \right)^2 \tag{2.4}$$

Equation 2.4 demonstrates higher frequency waves will encounter higher losses. Therefore, frequency can be used to classify the different EM waves and different radar types.

The different radar bands highlight the different applications for each type. VHF band radars (30-300 MHz [11]) will have lower propagation losses due to a lower frequency, and thus can be used in ground-penetrating applications. But the lower frequency and larger wavelength means the antenna will need to be larger. In contrast, an X-band (8-12 GHz [11]) radar will have high propagation losses, but allows for a smaller antenna, and offers capability of producing high-resolution images.

2.3 Antennas

An antenna is a transducer that is able to convert electromagnetic energy in the form of an electric current to a wave propagating in space (or any other material), or convert a wave in space back to an electric current. Antennas are fundamental because they enable radars to sense targets or its surrounding environment. A radar transmitter will generate a signal (in the form of an electric current), which then passes through RF hardware and an amplifier before reaching the antenna. The antenna then converts the waveform/signal into a propagating wave. The transmitted wave could possibly encounter a target, which will cause the wave to reflect and then be received by the same antenna (in the case of a monostatic setup), or another antenna (in the case of a bistatic or MIMO setup). At this point, the antenna and radar will be in receiving mode, "looking" for a waveform similar to the one transmitted albeit with considerable attenuation. The antenna in receive mode will convert the return wave into a signal/electric current. The signal is then processed to extract information about the target and environment e.g. target speed, range to target. Antennas can be constructed in various ways for various purposes. Some examples of antenna geometries include: parabolic reflectors, and phased arrays.

Phased array antennas have several advantages including: high bandwidth, high reliability, excellent sidelobe control, no moving or rotating parts and therefore excellent for stealth applications and for minimizing aircraft drag, and ideal for ground applications where rotation is impractical [11]. Unfortunately, much of this additional capability comes at a higher financial cost [11].

2.4 Waveforms

Radar waveforms come in two main classes: continuous-wave (CW) and pulsed. For CW radars, the transmitter and receiver operate simultaneously, but in order to prevent the transmit signal from damaging the receiver (due to proximity), the transmit power is less than that of a pulsed radar. This in turn limits the usable range of a CW radar. A CW radar is able to measure the Doppler shift on the return signal which can be used to determine the target's velocity. Since a CW radar is always transmitting, determining the target's range is slightly complicated: the signal's frequency changes over time (frequency modulation), which effectively provides timestamps, allowing the target's range to be determined [11]. One common application of CW radar is police speed radar [11].

Pulsed radars transmit bursts of EM energy on short timescales, typically on the order of microseconds, but could be as much as milliseconds or as little as nanoseconds. When the transmitter is on, the receiver is switched off to protect the hardware. Once an entire pulse is transmitted, the transmitter is switched off and the receiver is switched on so it can "listen" for the target echoes. Once the echo is received, the radar can begin processing it to learn more about the target and environment.

2.5 Measured Parameters

Knowledge about the beam characteristics and waveform as well as information gleaned from target echoes allow a radar to determine the following parameters of the target

- Azimuth angle, θ;
- Elevation angle, ϕ;
- Range, R and;
- Target velocity, v_r.

The target's angular position can be determined from the location of the antenna's main beam as it tracks the target [11]. The target's range is determined from the propagation time between the transmitted pulse and received echo. If the radar measures ΔT seconds from the time a pulse was transmitted to when the echo was received, then the target's range is

$$R = \frac{c\Delta T}{2} \,. \tag{2.5}$$

If the target is in motion, it will impart a Doppler shift, f_d onto the carrier frequency. The receiver will detect this shift and use it to determine target radial velocity as

$$v_r \approx \frac{f_d \lambda}{2}\,. \tag{2.6}$$

Pulsed waveforms' time domain characteristics are defined by the following (not exhaustive): (1) Pulse Repetition Frequency, PRF, (2) Pulse Width, τ. Pulse repetition frequency is how often pulses are transmitted, and pulse width is the amount of time a pulse is on.

The pulse width defines the range resolution, i.e. how large or small a range cell is. Smaller values of τ result in better range resolutions. The pulse repetition frequency defines the unambiguous range and the maximum detectable Doppler shift. The unambiguous range is given by

$$R_{ua} = \frac{c}{2 \cdot PRF} \tag{2.7}$$

and the maximum detectable Doppler shift is given by

$$f_{d_{max}} = \pm\frac{PRF}{2}\,. \tag{2.8}$$

The unambiguous range is the maximum range at which a target's range returns the correct value. Targets that lie beyond the unambiguous range will have their range values aliased and will appear closer to the radar than they actually are. The maximum Doppler shift is the highest frequency shift, and in turn, the highest permissible target velocity. If the target's Doppler shift is higher than this limit, it will be aliased (note that this result is related to Nyquist's Sampling Theorem). The unambiguous range and maximum Doppler shift produce a conflict because higher PRFs provide smaller range resolutions, but allow for higher maximum Doppler shifts. Conversely, lower PRFs produce higher unambiguous range, but lower maximum Doppler shifts. Therefore the selection of PRF (as well as other parameters) will be greatly influenced by the application, e.g. tracking long range targets will motivate lower PRFs, whereas tracking high-speed military aircraft will motivate higher PRFs.

There are techniques that can be employed to improve radar performance: pulse compression, linear frequency modulation, and biphase coding [11]. Pulse compression was developed to resolve the conflict between pulse energy and range resolution. Increasing pulse width increases energy but degrades the range resolution and vice-versa. Pulse compression decouples this relationship between pulse energy and range resolution, such that bandwidth can be increased without decreasing the pulse length. Today, linear frequency modulation and phase-coded waveforms are two techniques used to achieve pulse compression.

Linear frequency modulation (LFM) is based on a sinusoid whose frequency varies linearly with time. It has some unique properties that include Doppler tolerance (degree of degradation due to uncompensated Doppler), and is employed in radar systems supporting search, track, and high resolution modes [11].

Phase-coded waveforms are composed of concatenated subpulses (or chips) where the phase sequencing/coding/modulation from subpulse to subpulse is chosen to elicit desired time-domain mainlobe and sidelobe characteristics of the matched-filter response [11]. Some polyphase codes are Doppler tolerant, but others like biphase codes are Doppler intolerant when the Doppler shift exceeds one-quarter cycle over the uncompressed pulse length [11].

At the center of radar engineering is the radar range equation - an extension of the Friis transmission equation. Assuming a radar with one antenna for transmit and one for receive, where both antennas are co-located [11]:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} \tag{2.9}$$

where

P_r = Received power (W)

P_t = Transmitted power (W)

G_t = Transmit antenna gain (unitless)

G_r = Receive antenna gain (unitless)

λ = Wavelength (m)

σ = Target radar cross section (m^2)

R = Range to target (m).

Since $G_t = G_r = G$, we have

$$P_r = \frac{P_t G^2 \lambda^2 \sigma}{(4\pi)^3 R^4}\,. \tag{2.10}$$

The equation can be extended to account for bi-static cases, in which the gain of each antenna and the range to each antenna is considered [11]

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 {R_t}^2 {R_r}^2} \tag{2.11}$$

where

R_t = Range from transmitting antenna to target (m)

R_r = Range from target to receiving antenna (m).

The radar range equation also has the flexibility to account for noise. Assuming additive white Gaussian noise [11]

$$P_n = kT_sB = kT_0FB \tag{2.12}$$

where

P_n = Noise power (W)

k = Boltzmann's constant (J/K)

T_s = System noise temperature (K)

T_0 = Standard room temperature (290 K)

F = Noise factor (unitless; noise figure NF is the decibel verion of noise factor)

B = Instantaneous system bandwidth (Hz)

then Equation 2.9 can be used to determine the SNR of the received signal as [11]

$$SNR = \frac{P_r}{P_n} = \frac{P_t G^2 \lambda^2 \sigma}{(4\pi)^3 kT_0FBR^4}\,. \tag{2.13}$$

The radar range equation can also account for signal processing gains. Instead of detecting a single pulse, the radar can coherently integrate multiple pulses. If N_p pulses are integrated, then the SNR will improve by a factor of N_p [11]:

$$SNR = \frac{P_t G^2 \lambda^2 \sigma N_p}{(4\pi)^3 \, k T_0 F B R^4} \, . \tag{2.14}$$

Since systems are not ideal, the radar range equation should also account for losses, which can come in different types: transmit loss, atmospheric loss, receive loss, signal processing loss. The losses can be cumulatively described as system losses, defined as [11]

$$L_s = L_t L_a L_r L_{sp} \tag{2.15}$$

where

L_s = System loss (unitless)

L_t = Transmit loss (unitless)

L_a = Atmospheric loss (unitless)

L_r = Receive loss (unitless)

L_{sp} = Signal processing loss (unitless)

can be incorporated into the radar range equation as

$$SNR = \frac{P_t G^2 \lambda^2 \sigma N_p}{(4\pi)^3 \, k T_0 F B R^4 L_s} \, . \tag{2.16}$$

The radar range equation also allows for other variables to be solved for, namely the detectable range for a target with a given SNR and RCS; or for minimum RCS for a target at a given range and SNR [11]:

$$R_{det} = \left[\frac{P_t G^2 \lambda^2 \sigma N_p}{(4\pi)^3 \, k T_0 F B L_s \cdot SNR} \right]^{1/4} , \tag{2.17}$$

$$\sigma_{min} = \frac{(4\pi)^3 \, k T_0 F B R^4 L_s \cdot SNR}{P_t G^2 \lambda^2 \sigma N_p} . \tag{2.18}$$

2.6 Radar Functions

There are three basic functions of radar: search/detect, track, and imaging [11].

2.6.1 Search/Target Detection

Nearly all radars search for and detect targets without *a priori* information about the targets' presence or position [11]. Mechanically-steered antennas sweep through the search volume continuously whereas electronically scanning/phased-array antennas point the main beam to a series of discrete positions. At each position, one or more pulses are transmitted and received echoes are processed to detect a target. In the case of multiple pulses per position, the received echoes are non-coherently integrated to improve the signal-to-noise ratio of the observed position. The integrated data is compared against a threshold to make a decision on whether or not a target exists. This procedure runs through the entire search volume before repeating.

2.6.2 Target Tracking

Once a target has been detected, a radar can begin to measure the target's state; its position in range, azimuth angle, elevation angle, and radial velocity [11]. The individual position measurements are combined and smoothed to estimate a target track. Improved estimates of target track are obtained using track filtering or Kalman filtering [11].

It is worth noting that sometimes search and tracking functions are not performed by the same physical radar. Searching will typically have a wider beamwidth than tracking functions. Often one radar is fine-tuned for searching and another fine-tuned for tracking [11]. These setups are more common on land and surface ship applications. However, this is not possible on airborne platforms where space and power are limited [11]. Therefore, aircraft utilize a single radar that is a design compromise between the ideal search radar and ideal tracking radar [11].

2.6.3 Imaging

Radar imaging involves two steps: (1) developing a high-resolution range profile (HRRP) of the target; and (2) developing a high resolution cross-range (angular) profile. An example of radar imaging is synthetic aperture radar (SAR). SAR develops finely detailed images from a aircraft or spacecraft platform and has uses in surveillance, mapping, and resource monitoring. SAR systems may also be involved with identification of the objects in the images [11], *e.g.*, identifying non-cooperative tanks and vehicles.

2.7 Radar Applications

Although radars are common in military applications, there are many other areas as well where radars can be applied. The following is a short list of those applications [11].

1. Military Applications

   (a) Search Radar: Detects targets in the environment.

   (b) Air Defense Systems: Detects, tracks, and identifies airborne threats.

   (c) Over-the-horizon Search Radar: Utilizes refractive effects of the ionosphere in HF band to detect targets beyond the line-of-sight or horizon for conventional radars.

   (d) Ballistic Missile Defense Radar: Searches a large volume and able to track low-RCS, fast-moving targets

   (e) Instrumentation/Track Test Range Radar: Utilizes large antennas to achieve narrow beamwidths and long dwell times to obtain accurate measurements of targets. Can also provide inverse SAR images to train pattern-recognition-based target identification systems.

2. Commercial Applications

   (a) Process Control Radar: A non-contact method of measuring the amount/level of fluid inside of a tank. Typically utilizing frequency modulated continuous wave

(FMCW) at higher frequencies (10 GHz) to measure the distance down to the top of the fluid.

(b) Airport Surveillance Radar: Detects and tracks commercial and general aviation aircraft. Typically rotate mechanically in azimuth and have wide elevation beamwidths. Used in conjunction with a transponder to report flight number and altitude back to surveillance radar.

(c) Weather Radar: Measures the reflectivity of precipitation to obtain rainfall rate, uses Doppler techniques to obtain wind speed, and spectral width to measure turbulence. Some weather radars can use polarization characteristics of precipitation to discriminate between rain and hail, and others use Doppler techniques to measure wind shear, and rotating atmospheric (tornadoes) events.

A related application is radio-acoustic sounding systems (RAAS). An acoustic wave is transmitted vertically, followed by a radio wave also vertically oriented. The compression of air molecules caused by the acoustic wave changes the dielectric properties of the air, and produces detectable Doppler shift in the radar backscatter. The speed of the wave can be obtained from the Doppler shift, and since temperature of air is related to acoustic speed, the temperature profile of the atmosphere can be inferred.

(d) Wake Turbulence Detection: Large, heavy aircraft generate wake vorticies and turbulence behind them, and thus pose danger to smaller, lighter aircraft. Aircraft taking off and landing are separated by certain amount of time to allow the turbulence to dissipate. Radars placed at the end of runways can sense this turbulence and generate a warning for dangerous conditions.

(e) Satellite Mapping Radars: Satellites have the advantage of an unobstructed view of the Earth [11], and can operate at night or in poor weather conditions. Pulse compression techniques and SAR are used obtain good range and cross-range resolutions.

(f) Police Speed Radar: Utilizes continuous wave (CW) transmissions to measure the Doppler shift from a moving vehicle, which is then used to calculate the vehicle's speed.

(g) Automotive Collision Avoidance Radar: Currently deployed in some cars; utilizes a millimeter wave radar to scan the road for targets that may pose a risk of collision.

(h) Ground Penetration Radar: Utilizes a lower-frequency (L-band and lower) that can penetrate the ground and detect dielectric anomalies. Commonly used to detect buried pipes, gas leaks, buried land mines, tunnel detection, concrete evaluation and void detection in pavement.

(i) Radar Altimeter: Installed onboard aircraft and uses FMCW to measure the range to the ground, which will be the aircraft's height above ground.

Chapter 3

Introduction to Cognitive Radar and Machine Learning

3.1 Cognitive Radar Concept and Inspiration

Initially introduced by Haykin in his 2006 seminal paper [14], cognitive radar draws analogies from biological cognition. Cognition is defined as "knowing, perceiving, or conceiving as an act" [14, 15]. Humans perceive their environment through auditory and visual senses, process that information to learn more about the environment, and act on that information (*i.e.* make a decision).

There are animals, other than humans, that also demonstrate characteristics of cognition applicable to the work presented here. Bats, many of which are also blind, use sonar to navigate their environment and locate targets [5, 16]. Those bats that can echolocate have waveform characteristics that vary both with species and situation [16, 17]. As discussed in

[14], spectrograms of four different bat species illustrate how the repetition rate increases as the bat approaches its target. Over the course of their lives, these bats gained experience by trying different repetition rates, and use that experience to learn which rates to use (low rate versus high rate) when tracking a target [14].

Adaptive echolocation has also been noted in dolphins; the propagation of sound in water is superior to that of other forms of energy (e.g. light), thus making echolocation ideal for underwater navigation, object avoidance, and prey detection [18]. Target detection experiments with the *Tursiops truncatus* (Atlantic bottlenose dolphins) noted there was a corresponding increase in the number of transmitted *clicks* (analogous to radar pulses) to compensate for decreased SNR of echoes [18–21].

Electrolocation is a process used by weakly electric fish to navigate their surroundings [22]. These fish have an electric organ to generate an electric field around them, and surrounding objects that have a different electrical impedance compared to the water produce distortions in the field [22]. Electroreceptors on the body of the fish sense the distortions due to the presence of objects or the fields of other electric fish [23].The *Eigenmannia* (South American gymnotid), for example, continuosuly generates a quasi-sinusoidal discharge [23] of 1 V at 300 Hz [24]. When two electric fish encounter each other and have similar discharge frequencies, they risk jamming each other's electrolocation capabilities [23]. Some electric fish, like *Eigenmannia*, exhibit the *jamming avoidance response*, whereby each individual fish will shift their discharge frequency away (one will shift up, and the other will shift down) from the nominal frequency to minimize mutual jamming to their electrolocation senses [23].

As mentioned in Chapter 1, cognition has been built into radars in many ways. Cognitive radar models are able to perform a wide variety of functions such as adjusting the center frequency and bandwidth via optimization to mitigate the risk of interference [25–28], and adjusting pulse repetition rate to prevent a target from being Doppler aliased and being mapped into the Doppler clutter [29]. The field is not limited to these applications, however. Prior works in cognitive radar include applications to beamforming, target classification, waveform optimization and waveform diversity, target tracking, and spectrum sensing and spectrum agility.

3.2 Prior Work in Cognitive Radar

3.2.1 Beamforming

Basit et al. propose a beamforming technique for frequency diverse arrays that allow the radar to localize multiple targets in the same direction but with different ranges [30]. A frequency diverse array (FDA) is a generalization of phased-array radars, whereby each antenna component has a small frequency offset added to its carrier frequency [30]. The technique in [30] estimates a target's direction-of-arrival from the MUSIC algorithm and a target's range from the conventional range equation. The transmitter has a genetic algorithm which calculates a set non-uniform frequency offsets based on the future range and angle of the targets. The new frequency increments define the beam pattern for the next scan. New radar returns are received based on the new FDA beam pattern, and the above process repeats [30]. A genetic algorithm (GA) is a heuristic method based on biological evolution. It works by creating an initial set of random "chromosomes" where the chromosomes represent values that need to be optimized. The fitness of each chromosome is calculated, and then crossover is performed on the chromosomes by combining one chromosome with a different chromosome (akin to biological reproduction). Mutation is then performed on the offspring chromosomes. This process repeats until there is a chromosome that has the best available fitness, or the cycle limit of the algorithm is reached [30].

Sharaga et al. develop a beam pattern optimization technique for a MIMO Radar-Sonar system in an uncertain environment. The proposed target tracking algorithm is applied using sequential Bayesian filtering, implemented by particle filtering. The sequential conditional Bayesian Cramer-Rao Bound is chosen as the adaptive optimization criterion [31]. Particle filtering is a Monte Carlo methodology in which probability distributions are recursively approximated [32]. The Bayesian Cramer-Rao Bound provides a "tight and useful lower bound for estimation error [33]." Simulations demonstrated than even in an underwater environment with low SNR (0 dB), and there is considerable improvement over existing techniques, such as orthogonal beam forming [31].

3.2.2 Target Classification

Lunden and Koivunen develop a target recognition technique for multistatic radar systems [34]. High-resolution range profiles (HRRPs) are obtained by taking the inverse Fourier Transform of the far-field scattered electric field of a point-scatterer target. The HRRP profiles are normalized to the interval from 0 to 1 and fed to a convolutional neural network (CNN). The CNN's outputs are approximations of the target's posterior probabilities. Each radar system has a local classifier (the CNN mentioned above) and the outputs from each radar node are combined to form a global classification decision.

Lombacher et al. analyze the potential of radar for static object classification using deep learning methods [35]. Potential objects are extracted from an occupancy grid map via connected component analysis. Training data is selected by cutting a window around each object. The windows are also rotated from 0 to 360 degrees in 15 degree steps to account for various orientations. An equally distributed prior is assumed for all object classes because it is difficult to estimate a good prior distribution of the object's classes in the environment. This is achieved by oversampling the unbalanced set in two steps. The multi-class set is balanced so all classes are equally distributed, then the dataset is transformed into a one-vs-rest. The examined class is heavily oversampled. The analysis uses the CAFFE (Convolutional Architecture for Fast Feature Embedding) framework for neural network processing. The application for this technique would be for automotive radar.

Vasalos et al. outline a neural network target classifier for concealed weapon radar detectors [36]. The specific application involves using radars to detect and classify weapons, such as a gun, hidden on a person's body. The weapon and human body have specific resonant frequencies, called a Late Time Response in the literature, when separated, can enable target identification. For classification, the authors use a Learning Vector Quantization network. It is a neural network that combines a competitive layer and a linear layer.

Nijsure et al. discuss the application of an UWB MIMO radar onboard a UAV [37]. The radar mentioned in this paper utilizes a 2D-MUSIC algorithm for azimuth and elevation angle estimation. The Dirichlet-Process Mixture Model (DPMM) clustering framework is invoked to perform target detection and target discrimination. The DPMM provides a method of unsupervised mixture component analysis to discriminate between distinct UAV targets without a priori information about the target scene.

Bentes et al. present an application of neural networks to classifying oceanographic targets: cargo ships, tanker ships, oil platforms and wind farms, from synthetic aperture radar (SAR) images [38]. Prior neural network architectures for classification typically have a feed-forward, shallow architecture with an input layer, one hidden layer, and an output layer, combined with back-propagation and gradient-descent. Although they are able to solve complex problems in SAR image analysis, they are unable to take advantage of unlabeled data during the training process. In many cases, the input features need to be tuned to reduce the overall complexity. The authors of this paper present a deep neural network architecture that utilizes an autoencoder for each of the hidden layers. An autoencoder is a special configuration of a neural network that takes advantage of unlabeled data to learn the underlying information structure by a latent representation known as a code. In their architecture, a SAR image passes through a CFAR detector, which builds a list of detection targets. Each detection target defines a sub-image region of interest, and each image is pre-processed, filtered and re-scaled. The deep neural network consists of an unsupervised-trained block and a supervised-trained layer. The unsupervised block consists of a set of autoencoders and the supervised layer is trained on human-labeled data contained in the form of a database. The paper is only an extended abstract; it does not present simulation results and analyses.

Chen et al. present an application of deep convolutional neural networks to classifying SAR images [39]. Convolutional neural networks have achieved state-of-the-art results in computer vision applications, but have sever overfitting issues when directly applied to SAR images. This is dues to an insufficient number of training images available and an excess of free parameters. The authors propose a technique (all-convolutional NN, or A-ConvNets) that reduces the number of free parameters by utilizing sparsely-connected layers instead of fully connected layers. When evaluated with the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset, the algorithm is able to achieve 99% accuracy under standard operating conditions, and at least 96% under extended operating conditions (e.g. more variation in depression angle), and outperforms all other classification techniques they tested against, which include: EMACH, SVM, AdaBoost, Conditional Gaussian, IGT, MSRC, MSS, and M-PMC.

Scherreik and Rigling present a classification technique that deals with unlabeled data [40]. Many current classification problems involve closed sets, where of the classes that could possibly be detected are presented to the machine learning algorithm during training. To evaluate the algorithm's performance, samples are subjected to noise or some other perturbation or distortion. When a algorithm trained on a closed set is presented with a class it has not seen before, it gives labels that are often incorrect. The authors present their solution to this problem, called Probabilistic Open Set SVM (POS-SVM), which is an open-set recognition technique. Open-set recognition algorithms solve the aforementioned problem by having the option to forgo making a decision on an input that was not seen during training. This does not necessarily mean the input is discarded; it can be passed along to another algorithm (e.g. for online learning), or utilized in a human-in-the-loop system.

Benedetto et al. present a automatic aircraft target recognition technique based on processing of inverse-SAR (ISAR) images [41]. Inverse SAR, as opposed to conventional SAR, has a stationary radar platform and uses the motion of the target to produce an image of it. The ISAR images are processed by removing speckle noise via a linear filter followed by a median filter. The images are then segmented via the Smallest Univalue Segment Assimilating Nucleus (SUSAN) method, then Distance Regularized Level Set Evolution (DRLSE) is utilized to extract the target shape's contour. Once the target aircraft's contour is determined, Fourier Descriptors are used for feature extraction. Fourier Descriptors map each pixel in an image to frequency content. Using only the low-frequency content allows the generalized shape of the object to be reconstructed, while using all of the frequency content allows for the object to be fully reconstructed. Fourier descriptors are "useful for recognition tasks because [they] can be designed to be independent of scaling, translation, or rotation [42]." Fourier descriptors produce a vector of 168 samples, which are input into the neural classifier. The proposed algorithm classifies at 81.60%, and performs better than k-NN and SVM. Future work will consist of improving the individual neural networks, applying new search algorithms to improve generalization of neural networks, and improved image processing algorithms by going off other concepts in the literature.

Martorella et al. propose a technique of identifying targets from Polarimetric ISAR images [43]. The feature extraction process involves extracting the brightest scatterers using the Pol-CLEAN algorithm. The algorithm works iteratively by locating the brightest scatterer and finding its corresponding coordinate in the delay-Doppler domain; estimating target motion parameters and its point-spread function (PSF); and removing the scatterer from the Pol-ISAR image to find the next brightest scatterer. Once the scatterers are extracted, they are characterized according to Cameron's decomposition, which is a feature reduction technique. A single scattering matrix can be reduced to three variables; A set of N matrices will be reduced to $3N$ features, which will be the input size of the neural network. The Neural classifier is a multilayer perceptron (MLP), utilizing Marquardt backpropagation for training. The hidden neurons use sigmoidal activation functions and the output layers use linear activation functions. One advantage to using Polarimetric ISAR is the independence on the rotation of the target in the image; however the Pol-CLEAN method is disadvantaged by its high computational load.

Kim et al. present a target recognition technique using the MUSIC algorithm [44]. MUSIC generates one-dimensional range profiles, then central moments are calculated to provide translation-invariant and level-invariant feature sets. Principal Component Analysis is then conducted to reduce the feature set size. Finally, the reduced feature set is input to a Bayes classifier for recognition. The MUSIC algorithm is shown to produce range profiles that in turn, have higher correct classification results than the IFFT.

3.2.3 Waveform Optimization and Waveform Diversity

There are many developments in cognitive radar with respect to waveform optimization and waveform diversity. Zhang et al. propose a waveform selection technique based on what they call the "wind-driven optimization technique" [45]. Wind-driven optimization technique is based on the physical motion of particles in windy conditions. It starts with a population of air parcels at random positions and with random velocities. On each iteration of the algorithm, each parcel of air's position and velocity are updated, and as time progresses the parcels will move toward an optimum solution at the end of the iterations. The authors of this paper propose using the wind-driven optimization technique to minimize the predicted tracking Cramer-Rao Lower Bound.

Rongwen et al. [46] propose a waveform selection method for anti-passive false target jammers. It uses the distinction degree as the criterion for selecting an optimal waveform to be used while a jammer is present in the environment. Chen and Wu [47] discuss a waveform design technique based on the water-filling algorithm to optimize the power spectral density (PSD) of the waveform for signal target detection.

La Manna et al. describe a spectrum-controlled waveform for use in a cognitive radar [48]. The implemented radar system has a cognitive optimizer on the receiver and another optimizer on the transmitter and proposed solution is called Adaptive Spectrum Controlled Waveform (ASCW). The transmitter implements frequency nulling on the waveform to reduce interference to co-existing communication signals. In addition, the receiver reduces interference to the radar due to other communication systems.

Yuang et al. [49] describe a waveform optimization for cognitive radars operating in environment with interference. The optimization technique invokes Wiener filtering theory and the Cauchy-Schwarz theorem to describe the optimal waveform in the presence of colored tones (e.g. jammers, interfering tones). One drawback to this technique is optimization requires prior knowledge of the jamming waveform. But obtaining this knowledge, which could be in the form of an autocorrelation matrix requires accumulating multiple echoes to improve the jamming estimate. But if the jammer is frequency agile, it will be very difficult to obtain the autocorrelation matrix estimate.

Martone et al. present the concept of cognitive nonlinear radar in [50]. A nonlinear radar differs from traditional radar in that the radar returns are not at the same frequency as the transmit waveform; this change in frequency is attributed to the characteristics of the target material. The radar presented in the report transmits waveforms in various bands, and senses for the returns in different bands. The cognitive nonlinear radar optimizes its waveform based on interference, target likelihood and permissible transmit frequencies as allowed by regulations and other users in the environment. A cognitive nonlinear radar could have many challenges and conflicting objectives; for example using optimal bands for detecting a target without interfering with other users. A set of objective functions are proposed, and optimization is performed to obtain optimal values.

3.2.4 Target Tracking

Martone et al. present a spectrum sensing technique that enables a cognitive radar to select an optimal sub-band that optimizes range resolution and signal-to-interference-and-noise ratio (SINR) [25]. Optimizing on range resolution and SINR are conflicting tasks because a better range resolution requires a wider bandwidth. However, a wider bandwidth introduces more noise to the receiver ($P = kTB$), therefore reducing the SINR. This conflict in objectives is resolved by developing one objective function for optimizing range resolution and one objective function for optimizing SINR. The two objective functions are combined using a linear-weighted multi-objective function. The output from the multi-objective function is an optimal value for bandwidth and the center frequency for the optimal band. The optimal bandwidth and center frequencies are fed to the transmitter to optimize the transmit waveform, and this process is repeated for each transmit/receive cycle. Future work on this topic includes reducing computational complexity of the algorithm and combining multiple, discontinuous sub-bands to maximize the available bandwidth for the radar to use.

Martone et al. in [51] present an application of the adaptable bandwidth selection algorithm from [25] to harmonic step frequency radar. Harmonic radars process radar echoes that are harmonics of the transmit frequency, which result from “nonlinear scattering by targets of interest.” The harmonic returns also appear in harmonic multiples of the transmit bandwidth, while clutter appears only in the same band as the transmit frequency [51]. This fact facilitates the detection of nonlinear targets. Simulations indicated SINR improved by over 25 dB when an optimal subband is selected in the presence of noise. The authors of

[51] do note the technique does sacrifice some range resolution, as a result of select a smaller bandwidth, which makes separating closely spaced targets more difficult.

Wang et al. present a cognitive target tracking method to improve SINR performance in a frequency-diverse array (FDA) radar [52]. The radar develops estimates of the range and direction-of-arrival of a target and feeds this information from the receiver to transmitter. The transmitter then uses this information to update the frequency offset which is used to control the beampattern of the FDA radar. Meanwhile, the radar uses the minimum variance distortionless response beamformer to minimize the interference-plus-noise power.

Wang presents a moving-target cognitive tracking radar implemented with a frequency-diverse array antenna (FDA) [53]. The different frequency offsets sent to the antenna elements not only create the FDA beampattern, but also reduce the peak power of the radar signal to make the energy at an unintended receiver difficult to detect. The author uses a quadratic phase slope across the array to reduce the antenna's gain, and the quadratic phase variation is calculated by a multidimensional gradient search routine. The transmitter calculates frequency offsets and phase offsets to create a beampattern, and the receiver analyzes the energy reflected off the target and performs target tracking. Then the radar receiver analyzes its performance in the context of SNR and the tracking results (range and angle), and via a feedback loop to the transmitter, these values will be used to adjust the transmit beampattern on the next scan. This application is a fore-active radar (FAR); while there is a feedback loop and processing is done on echoes from the previous cycle, it lacks aspects of intelligence that Haykin mentions is key to cognitive radar.

Kreucher et al. present a comparison of tracking algorithms for supermaneuverable aircraft targets [54]. Supermaneuverable targets are aircraft able to perform high-G maneuvers beyond the capabilities of most aircraft - typically military aircraft. The paper also considers aircraft with low-RCS. The algorithms of interest are the extended Kalman filter (EKF), the unscented Kalman filter (UKF), particle filter with resampling (PFR), and particle filter with homotopy flow (PFH). Results from simulations can be broadly summarized as follows: Kalman Filters are computationally efficient and work well with high-SNR, stable-RCS targets. Particle filters are more computationally expensive, but are able to more accurately model target motion uncertainty and work under low-RCS, high-scintillation, high-G conditions even when Kalman filters fail. The paper additionally notes that Kalman filters must detect the target before tracking it, whereas particle filters allow for track-before-detect approaches, which could propose an interesting avenue of research regarding detection and tracking of high-speed targets.

Bell et al. present a cognitive radar for tracking using a software-defined radar system [55]. The technique presented is based on the maximum a posteriori penalty function (MAP-PF) to obtain a track estimate of the target. The pulse-Doppler radar's controller adjusts the PRF to optimize the tracking performance. However, there are multiple conflicts associated with adjusting the PRF: (1) decreasing PRF results in increased uncertainty in the motion model; (2) as PRF decreases, the Doppler bin width decreases, which improves Doppler measurement resolution; (3) AS PRF decreases and Doppler bin width decreases, the target will be easier to discriminate from the bins with zero-Doppler clutter; and (4) As PRF decreases, the target will be Doppler aliased if the Doppler shift is greater than PRF/2. In their experiments, a human target moved back and forth in front of a radar, over a 5 meter span. As the target velocity peaked - when the target was in the midpoint of the span - the PRF was increased to its maximum value to prevent Doppler aliasing. When the velocity changed sign - when the target was either at the near or far ends of span and was changing direction - the PRF was decreased to enable easier target discrimination from the clutter. This application has a feedback loop, processes prior samples, and employs signal processing, but is ultimately adaptive; the radar doesn't learn from its prior experience. Thus, this is also a fore-active radar (FAR).

3.2.5 Spectrum Sensing and Spectrum Agility

Wabeke and Nel present an application of reinforcement learning to a frequency-agile radar adapting to its environment [56]. The radar presented in the paper is attempting to detect targets with varying scan lifetimes and incoming targets. The authors chose to implemented Q-Learning as the algorithm that selects the transmit frequency. Q-Learning is an efficient form of reinforcement learning for dynamic programming. Dynamic programming is a much older approach to determining optimal decision making policies for sequential optimization (the Viterbi decoder is an example of dynamic programming). The goal of Q-Learning is to choose an optimal policy at a given state that would correspond to choosing the action corresponding to the maximum value of Q in a particular state (Q represents the expected reward obtainable in a future state). In demonstrations, Q-Learning was shown to outperform other methods (random frequency selection, frequency sweeping and frequency hopping) all other methods in all cases except for the longest scan lifetimes because it has less frequency diversity than the frequency sweeping approach.

Oksanen et al. present a reinforcement-learning-based spectrum sensing approach in cognitive radio networks [57]. The network of cognitive radios can individually sense spectrum and report their findings to a fusion center that handles data processing. The network of radios frequency hop, utilizing pseudorandom orthogonal sequences to maximize the number of sensors covering as much of the spectrum as possible while minimizing the time spent sensing. The authors present a reinforcement learning algorithm called ε-greedy, which finds a balance between the time spent exploring (searching for bands) and exploiting (using a frequency band). Although the paper discusses an application for cognitive radios, particularly for battery-operated units, the same idea could apply for cognitive radios operating on mobile platforms such as an unmanned aerial vehicle (UAV), which has limited power source and whose spectral environment may change depending on location.

3.3 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) is a field of science that aims to understand and construct intelligent entities (machines) [9]. Definitions may vary, but [9] considers AI to be organized into any of the following definitions (1) Systems that think like humans, (2) Systems that act like humans, (3) Systems that think rationally[1], and (4) Systems that act rationally. Among applications of AI include the more general tasks such as learning and perception, to more specific tasks such as "playing chess, proving mathematical theorems, writing poetry, and diagnosing diseases [9]."

3.3.1 Reinforcement Learning

Reinforcement learning is concerned with using the concept of *reward* to serve as feedback on which actions are *good* and which ones are *bad*. This contrasts with other forms of machine learning such as supervised learning, in which a "teacher" acts as feedback, dictating which actions are good and bad. Reinforcement learning is useful in cases where it is impractical for a designer to manually provide information and evaluation about a large number of states [9]. Rather, the intelligent agent learns on its own which sequences of actions lead to more reward, and which ones will lead to less reward [9]. The goal behind reinforcement learning is to maximize the sum of reward; the optimal action or sequence of actions will return the highest amount of reward [9]. The reward provides a relative indication of quality of an action (desirable actions result in positive reward while undesirable actions result in negative reward). Part of the challenge of reinforcement learning is the environment information is not provided *a priori* [9]. The agent must *explore* its environment, learning which actions would be beneficial or detrimental [9].

[1]The authors of [9] define *rational* as an ideal concept of intelligence, or in other words "[A] system is rational if it does the "right thing", given what it knows." As the authors point out, *rational* does not suggest that humans are "irrational" in the sense of "emotionally unstable", but rather to acknowledge that humans are imperfect and can make errors in reasoning and logic. In contrast, a *rational* entity/system is not prone to errors in reasoning that a human could make.

3.3.2 Markov Decision Processes (MDPs)

Since the heart of our approach is MDPs, we first briefly describe them. MDPs are used to model planning for an autonomous agent in an uncertain environment [58]. MDPs are popular in two sub-fields within artificial intelligence, probabilistic planning and reinforcement learning [58]. The probabilistic planning literature focuses on developing computationally efficient approaches to solve MDPs, with the assumption that complete knowledge of the MDP is available [59]. Reinforcement learning however, is a more difficult problem in which the agent starts with no prior knowledge of the MDP and has to learn from experience by interacting and experimenting with its environment to gain knowledge about how to optimize its behavior [58, 59]. The work in this paper is of the reinforcement learning type in which our radar (the agent) learns characteristics of its environment through experience.

An MDP is specified by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, \pi^* \rangle$. $\mathcal{S}$ is the set of all possible states in the model, sometimes called the state space. A state $s \in \mathcal{S}$ is a unique characterization of environment information [59]. The action space $\mathcal{A}$ is the set of all actions that can be taken by the agent to control or change the state [59]. The transition probability function $\mathcal{T}(s, a, s')$, is a description of the probability that an agent in state $s \in \mathcal{S}$ will transition to another state $s' \in \mathcal{S}$ when taking action $a \in \mathcal{A}$. The Markovian attribute of MDPs means the future state as the result of an action does not depend on previous actions and states;

the future state only depends on the current state and current action, in other words [59]:

$$\begin{aligned} \mathbb{P}\left(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots\right) &= \mathbb{P}\left(s_{t+1} \mid s_t, a_t\right) \\ &= \mathcal{T}\left(s_t, a, s_{t+1}\right). \end{aligned} \tag{3.1}$$

Note that in our application, the transition function is assumed to be unknown in advance, and we use a frequentist approach to estimate it. The frequentist approach calculates the probability of an event ε via $\mathbb{P}\left(\varepsilon\right) = \lim\limits_{n\to\infty} \frac{n_\varepsilon}{n}$, where n_ε is the number of times event ε occurs, n is the total number of trials and the ratio n_ε/n is known as the relative frequency of event ε [60]. In our implementation, the probability is computed for each action a as such: $\mathcal{T}(s, a, s') = \mathbb{P}(s' \mid s) = N_{s'}/N_s$, where N_s is the number of times the agent is in state s, and $N_{s'}$ is the number of times the agent transitions to state s' from state s.

The reward function $\mathcal{R}\left(s, a, s'\right)$ is a description of the average reward accumulated by the agent when the agent was in state s, performed action a and transitioned to state s'. The values in the reward function could be positive (usual connotation of reward), or negative (punishment/penalty) [59]. Like the transition function, the reward function is unknown in advance and is estimated in the simulation.

The discount factor $\gamma \in [0, 1]$ models the preference for current rewards versus future rewards [9]. When γ is close to 0, the agent will prefer immediate rewards and future rewards will be heavily discounted [9]. When γ is close to 1, the agent will prefer the distant, long-term rewards. Discounting is a good model of animal and human behavior [9] and helps ensure that the utility of a state sequence is finite.

A value function (also known as utility)[2], in Equation 3.2, can be used to describe "*how good* it is for the agent to be in a certain state", given a particular policy π [59]:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| \pi, s_t = s\right]. \tag{3.2}$$

Following the development in [59], the value function can be expanded to Equation 3.7, where the value function $V^{\pi}(s)$ for the current state s, and given any policy π can be described in terms of the value function for the future state s', discount factor γ, and the transition probabilities $\mathcal{T}$ [59]. Equation 3.7 is also known as the Bellman Equation [59].

$$\begin{aligned}
V^{\pi}(s) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| \pi, s_t = s\right] && (3.3)\\
&= \mathbb{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \middle| \pi, s_t = s\right] && (3.4)\\
&= \mathbb{E}\left[R_t + \sum_{k=1}^{\infty} \gamma^k R_{t+k} \middle| \pi, s_t = s\right] && (3.5)\\
&= \mathbb{E}\left[R_t + \gamma V^{\pi}(s_{t+1}) \middle| \pi, s_t = s\right] && (3.6)\\
V^{\pi}(s) &= \sum_{s'} \mathcal{T}(s, a, s') \Big(\mathcal{R}(s, a, s') + \gamma V^{\pi}(s')\Big)\Big|_{a=\pi(s)} && (3.7)
\end{aligned}$$

The optimal policy π^* will be the one that results in the agent receiving the most reward, such that its value function is greater than that of any other possible realisation, or in other words $V^{\pi^*}(s) \geq V^{\pi}(s)\ \forall_{\pi,s}$ [59]. The value function for the optimal policy is defined and

[2] The term "utility" used in [9] is equivalent to the term "value function" used in [59]. Therefore, $U(s)$ used in [9] and $V(s)$ used in [59] are equivalent to each other.

known as the Bellman optimality equation [59]:

$$V^{\pi^*}(s) = V^*(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') \Big(\mathcal{R}(s, a, s') + \gamma V^{\pi}(s') \Big). \tag{3.8}$$

From which, the optimal policy is derived as [59]:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') \Big(\mathcal{R}(s, a, s') + \gamma V^{\pi}(s') \Big). \tag{3.9}$$

It is worth noting that in drawing connections between cognitive neuroscience and cognitive systems in [61], Haykin and Fuster link Bellman's dynamic programming as "the mathematical basis for cognitive control."

There are two primary methods for calculating the optimal policy, value iteration and policy iteration; the work presented in this paper uses policy iteration. The solver used is from MDPToolbox, a MATLAB toolbox developed by researchers from INRA Toulouse [62]. Policy iteration begins from some initial policy π_0 and alternates between two steps: policy evaluation, and policy improvement [9]. Policy evaluation calculates the utility of all states, given a policy π [9]:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| \pi, s_t = s \right]. \tag{3.10}$$

Policy improvement then uses the utility function $V^{\pi}(s)$ to choose the action a for the current state that maximizes the expected utility of the subsequent state s' [9]; thereby creating an updated policy π' [59]:

$$\pi'(s) = \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{T}\left(s, a, s'\right) V^{\pi}(s'). \tag{3.11}$$

Then the new policy π' is used to compute a new value function $V^{\pi'}$ (via policy evaluation), the result of which is used to create a newer policy (via policy improvement) [59]. This process repeats until the policy can no longer be improved, meaning the optimal policy π^* has been obtained [59].

3.3.3 Summary

Cognitive radar has a rich amount of research, covering fields from beamforming and target tracking, to target tracking and spectrum sensing/agility. However, there is a relative lack of research in the combination of target tracking and spectrum agility. The focus of this work extends the work in [27], and use Markov decision processes and reinforcement learning in place of on-line multi-objective optimization.

There are some works that involve applying MDPs to radar problems. These include resource management for airborne radar [63, 64], optimal sensor scheduling while tracking multiple targets [65], waveform selection for target detection [66], and adaptive beam scheduling for target tracking [67].

Chapter 4

System Model and Detailed Approach

4.1 Proposed System Model

The focus of this paper is applying the MDP framework to the radar tracking problem. To prevent the state space from becoming intractably large, we make simplifying assumptions about the radar scene. The target is a simple point target and is moving generally orthogonal to the boresight direction of the radar, although the exact trajectory on each training run is random (see Figure 4.2). The interferer is a communications system that can occupy one or more bands at a time, is physically stationary, and (except for the direction-dependent interferer) location independent (i.e. neither the interferer nor the target's position with respect to the radar affects the interference sensed by the radar). The environment is simple such that clutter is negligible and the radar returns are not subject to multipath or atmospheric effects (e.g. rain) other than the free space path loss given by the radar range equation. The radar uses a linear frequency modulated (LFM) chirp waveform with the appropriate time bandwidth product. Also, the radar can perfectly determine Doppler shift and target velocity, and use that perfect knowledge to account for the range-Doppler coupling effect as a result of using the LFM waveform.

4.2 The Radar Environment

An example radar scene is shown in Figure 4.1. The red circles represent position states (cells), and the blue line an example target trajectory. The radar environment is defined by a set of possible position states $\mathcal{X}$, and a set of possible velocity states $\mathcal{V}$,

$$\mathcal{X} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_\rho\}^T \tag{4.1}$$

$$\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\nu\}^T \tag{4.2}$$

where ρ is the number of possible positions, ν is the number of possible velocities, and T denotes the transpose operation. Each of the $\mathbf{r}_i$ is a 1×3 vector defined as

$$\mathbf{r}_i = [r_x, r_y, r_z] \tag{4.3}$$

where r_x, r_y, r_z are the position components in the cross-range, down-range, and vertical dimensions, respectively. Like the positions, each of the $\mathbf{v}_i$ is a 1×3 vector defined as

$$\mathbf{v}_i = [v_x, v_y, v_z] \tag{4.4}$$

where v_x, v_y, v_z are the velocity components. The radar is located at the origin, $(0, 0, 0)$. Note the plot is a top-down view of the radar scene, and therefore the vertical dimension is not shown.

Figure 4.1: An example radar scene and trajectory.

The interference states $\mathbf{\Theta}$ are defined as

$$\mathbf{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M\}^T \tag{4.5}$$

where M is the number of unique interference states. Given N frequency bands, the number of unique interference states is $M = 2^N$. Each of the $\boldsymbol{\theta}_i$ is a $1 \times N$ vector defined as

$$\boldsymbol{\theta}_i = [\theta_1,\, \theta_2,\, \ldots,\, \theta_N] \tag{4.6}$$

where the $\theta_i \in \{0, 1\}$ indicates if an interferer exists in the ith band. As an example, $\boldsymbol{\theta} = [0\,1\,0\,1]$ means there are 4 bands, of which the 2nd and 4th bands have interference present.

For our model, the set $\mathcal{S}$ denotes all the combinations of target position states, target velocity states, and interference states. The total number of states is $N_S = \rho\nu 2^N$. The actions $\mathcal{A}$ are defined as

$$\mathcal{A} = \{\boldsymbol{a}_1,\, \boldsymbol{a}_2,\, \ldots,\, \boldsymbol{a}_{N_A}\}^T \tag{4.7}$$

where N_A is the number of actions. Each of the $\boldsymbol{a}_i$ is a $1 \times N$ vector defined as

$$\boldsymbol{a}_i = [\alpha_1, \alpha_2, \ldots, \alpha_N] \tag{4.8}$$

where the $\alpha_i \in \{0, 1\}$ indicate whether or not the radar has selected a particular band in which to transmit its waveform. For example, $\boldsymbol{a} = [1\,1\,1\,0]$ means there are 4 bands, and the lowest three bands are used by the radar. Valid actions are those that use only contiguous groups of bands. Examples of valid actions include $[0\,0\,0\,1]$, $[0\,1\,1\,0]$, $[1\,1\,1\,1]$, but $[1\,0\,0\,1]$ and $[1\,1\,0\,1]$ are not valid actions. It can be shown that the number of valid actions is $N_A = [N(N+1)]/2$.

The transition probability function is defined as follows: $\mathcal{T}(s, a, s') : N_S \times N_A \times N_S \to [0, 1]$, where the first dimension represents the current state and the third dimension represents the future state, and all of its values are bounded on $[0, 1]$. Similarly, the reward function is defined as $\mathcal{R}(s, a, s') : N_S \times N_A \times N_S \to \mathbb{R}$, where its values are real numbers. On each iteration of the simulation, after the future state s_{t+1} is determined, the reward for that state $R(s_{t+1})$ based on action a_t is computed. The instantaneous reward is determined from the reward structure, which considers SINR and amount of bandwidth used by the radar. Note that reward is based on current conditions, whereas actions are decided based on immediately preceding conditions. The reward structure provides positive reward for higher SINR (up to some maximum value) and increased bandwidth usage, while penalizing negative SINR.

At the heart of this problem is the radar's range resolution, defined as

$$\Delta R = \frac{c}{2\beta} \tag{4.9}$$

where c is the speed of light, and β is the radar's bandwidth. Range resolution dictates the accuracy of the range measurement. When the target is further away, a coarse range resolution is acceptable. However, when the target approaches the radar, a coarse range resolution will produce an inaccurate range measurement. Finer range resolution is obtained by increasing the radar's bandwidth. However, if the radar also needs to coexist with a communications system in the same spectrum, there is a possibility for the radar to use the same bands as the communications system. In doing so, the radar uses the same band as the communications system (resulting in interference), which causes the SINR to drop. If the SINR drops sufficiently, the radar could lose the target, which is very undesirable for a tracking radar. There is therefore a conflict between range resolution, and SINR, both of which are linked by bandwidth. The goal of this work is to apply reinforcement learning technique to enable the radar to achieve optimal performance; to have fine range resolution by using as much bandwidth, while also mitigating interference to maintain positive SINR.

4.3 Experiment Details

The experiments involves two major steps: 1. Training, and 2. Testing. Training involves running the radar against scenarios that it may encounter. Many training runs (on the order of 10^3 to 10^5 depending on interference type) are needed. Each run is set up by selecting, at random, one position state, and one velocity state. Normally-distributed random "noise" is added to both the position and velocity, to ensure each trajectory is unique. A sample of random trajectories used for training is illustrated in Figure 4.2. During each training run, the current state s is determined, then a valid action a is selected at random. This is generally termed "exploration" in reinforcement learning. The amount of bandwidth is determined based on the action and the resulting interference based on the action and interference behavior is updated. The position and range are updated, and the resulting SINR is calculated. The new/future state s' (given the new interference and position) is determined, and the probability transition function $\mathcal{T}(s, a, s')$ and reward function $\mathcal{R}(s, a, s')$ are updated. When all of the training runs are complete, policy iteration uses the discount factor, the estimated probability transition function, and estimated reward function to compute the optimal policy. Then, we test the policy to see how well the radar has learned from its training. This is generally termed "exploitation" in reinforcement learning. Testing starts with a user-defined trajectory, which will be different than any of the trajectories the radar trained on. This demonstrates that the radar is able to generalize, and is not overtrained on any set of trajectories. Given the user-defined trajectory, the initial state s is determined, which is used to select an action from the policy; in other words $a = \pi^*(s)$. The bandwidth is computed from the action (which is given by the policy), and the the interference, position, range, and SINR are updated, as well as the resulting reward. The simulation is described in algorithmic form in Appendix A. The results below are based on testing the radar on the user-defined trajectory (*i.e.*, after training).

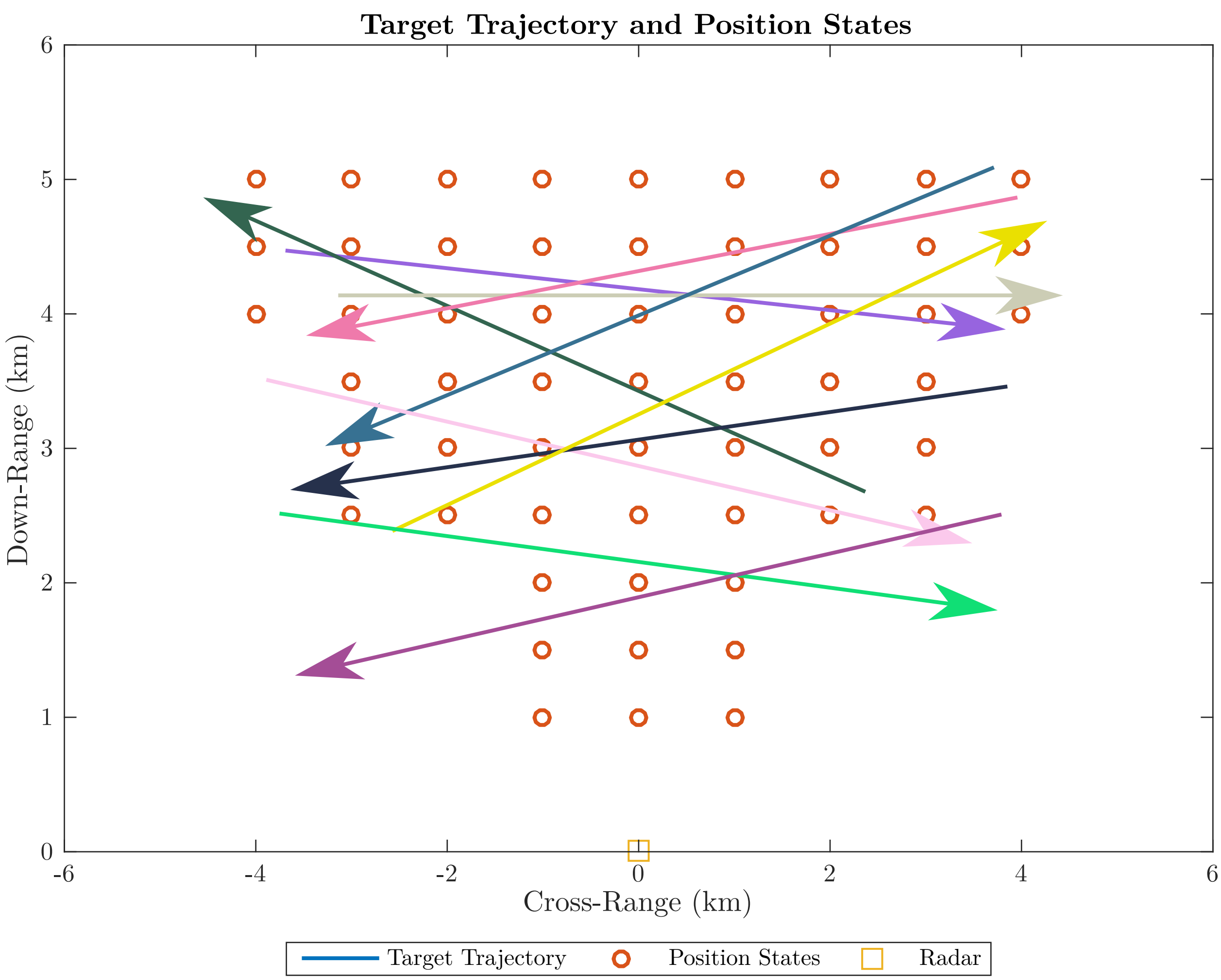

Figure 4.2: Example of the random trajectories used for training.

Chapter 5

Experimental Results and Analysis

The following results show the performance of the radar for each interference type. For each figure, the upper plot shows the cumulative rewards, the amount of bandwidth used by the radar, the target's range and the target's SINR over time. The rewards and bandwidth are plotted vs. the left y-axis, and the range and SINR are plotted vs. the right y-axis. The lower plot shows the interference and the actions taken by the radar. The numbers on the y-axis of the lower plot are the decimal conversions of $\boldsymbol{\theta}_i$ and $\boldsymbol{a}_i$; where $\boldsymbol{\theta}_i$ and $\boldsymbol{a}_i$ can be treated as vectors of binary values. For example, if the interference's action value is 16, that means the interference occupancy vector is $\boldsymbol{\theta}_i = [1\,0\,0\,0\,0]$, and if the radar's action number is 31, the radar occupancy vector is $\boldsymbol{a}_i = [1\,1\,1\,1\,1]$.

The reward structure greatly influences the behavior of the radar. In our experiments, the reward structure is set up such that if the SINR is negative and not all bands are used by the radar, the agent will receive a large net negative penalty. A negative penalty reflects the high probability of losing the target at negative SINR. When the SINR is negative, but all of the bands are used by the radar, the agent receives a small net positive reward; where the reward for using all bands is greater than the penalty for negative SINR. This reward structure provides some incentive for the radar to take some chances and use all of the bands, even if there is risk of having negative SINR. If the reward structure is changed to make the penalty for negative SINR greater than the reward for using all bands, the radar will be more conservative in its decision making and not take the risk of having a negative SINR. This could also be used to make the radar less likely to cause interference to communication systems. Overall, the radar's performance is dictated by SINR and bandwidth; multiplicative increases in bandwidth are more important than incremental increases in SINR. The reward for SINR also saturates at 20 dB to reflect that there is no practical benefit gained from having an SINR higher than some threshold. The reward structure with $N = 5$ bands (the value used in the simulations) is summarized in Table 5.1. The total reward at one time instant is determined from the sum of values from both columns. For example, if SINR = 3 dB, and the radar uses four bands, then the total reward at that time would be 2 + 30 = 32.

5.1 Constant interference

In the case of constant interference, the communications system occupies a non-zero number of bands and does not change for the duration of a training run. The motivation for this case is to test the performance of the MDP model against only the target trajectory. An example

Table 5.1: Summary of reward structure.

Summary of Reward Structure For $N = 5$ Bands			
SINR (dB)	Reward	Number of Bands Used	Reward
< 0	-35	1	$+0$
$0 - 2$	$+1$	2	$+10$
$2 - 5$	$+2$	3	$+20$
$5 - 8$	$+3$	4	$+30$
$8 - 11$	$+4$	5	$+40$
$11 - 14$	$+5$		
$14 - 17$	$+6$		
$17 - 20$	$+8$		
> 20	$+10$		

result for $\boldsymbol{\theta} = [1\,0\,0\,0\,0]$ is shown in shown in Figure 5.1. This example demonstrates that when the target is farther away, the radar avoids the interference by selecting all the bands where the interference does not exist. As the target crosses the radar environment, its range decreases and as a result, the SINR increases. When the SINR is sufficiently high, the radar can accept trading off SINR to use more bandwidth. At that point the radar is able to use all of the bands, even if one is occupied by the interferer, and the SINR is still positive. This behavior is a result of the reward structure. After the target makes its closest approach to the radar, its range begins to increase. When the target is sufficiently far away, the radar needs the SINR to stay positive and thus trades some of the bandwidth to improve SINR, using the same bands as in the beginning of the trajectory.

In this example, since the interferer occupies only a single band on the edge, the radar learns to occupy the remaining contiguous bands when the target is farther away since using all bands would drive the SINR negative. As the target comes closer to the radar, the received signal is strong enough to provide good SINR even in the presence of interference. Thus, the radar learns to use the entire band, reaping the benefit of larger bandwidth.

The rewards for this example are worked out in Tables 5.2 and 5.3 to demonstrate the optimality of the learned behavior. The rewards are computed for each possible action, when the target is farther away and when the target is closer to the radar. When the target is farther away (5.5 km), the action $\boldsymbol{a} = [0\,1\,1\,1\,1]$ returns the highest reward (35) because it uses the highest number of bands, while also keeping the SINR positive. Since this is the highest reward the radar can get, $\boldsymbol{a} = [0\,1\,1\,1\,1]$ is the optimal action *under those circumstances* (target is 5.5 km away). When the target is closer to the radar (*e.g.* 3.8 km), the SINR is sufficiently high to allow for using all of the bands. When the radar takes the action $\boldsymbol{a} = [1\,1\,1\,1\,1]$, it maximizes its reward by using all of the bands, as seen in Table 5.3. The additional reward due to bandwidth offsets the decrease in reward if the radar were to use fewer bands, but have a higher SINR by avoiding the interferer (compare action $[0\,1\,1\,1\,1]$ to $[1\,1\,1\,1\,1]$, where reward equals 38 and 42 respectively). Since 42 is the highest reward the radar can receive when the target is closer, $\boldsymbol{a} = [1\,1\,1\,1\,1]$ is the optimal action when the target is closer to the radar. We see this behavior in Figure 5.1.

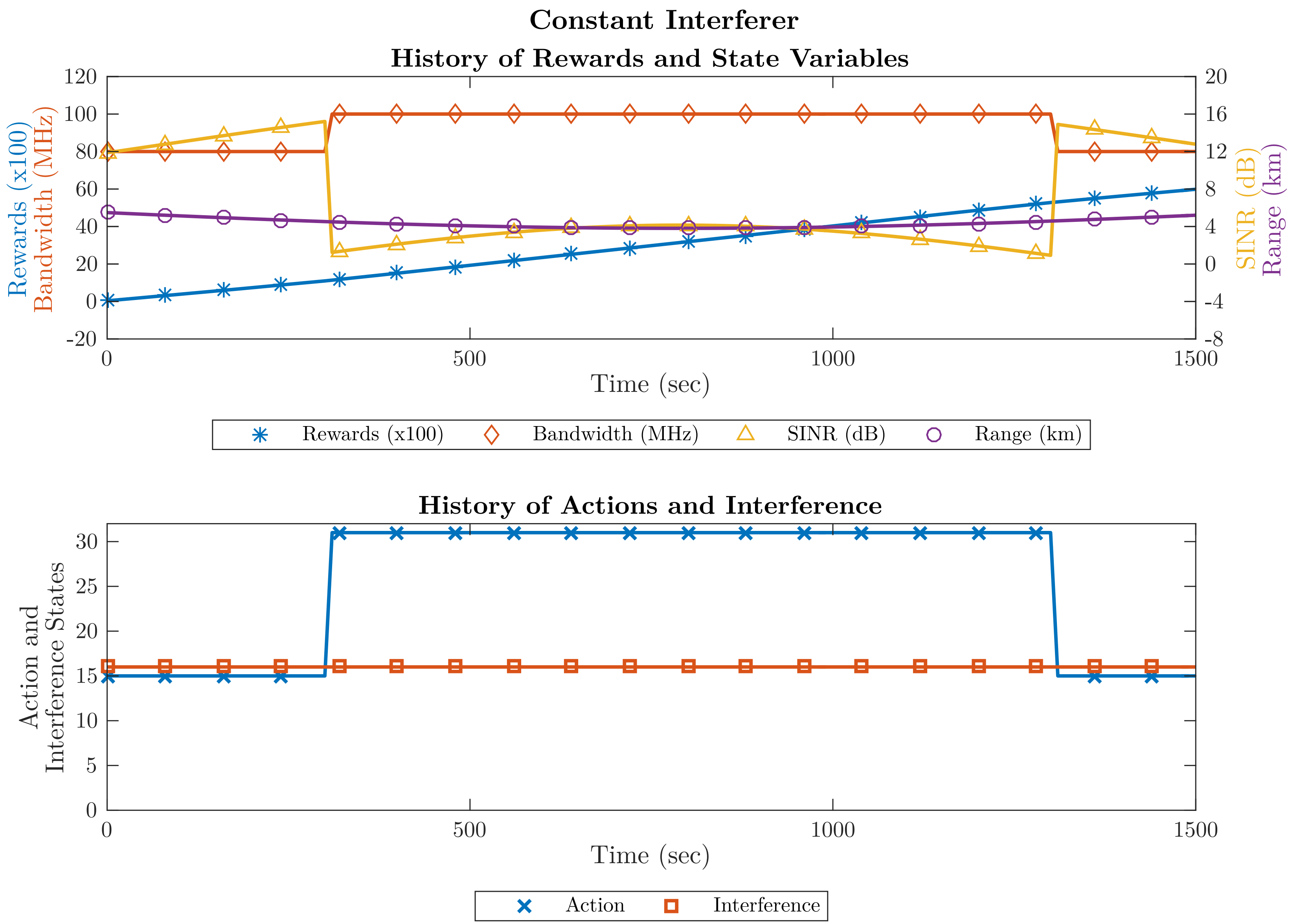

Figure 5.1: Results for constant interferer.

Table 5.2: Rewards for each action, when the target is further away from the radar.

Observed Interference	Range (km)	Action	BW (MHz)	SINR (dB)	Reward
[1 0 0 0 0]	5.5	[0 0 0 0 1]	20	11.8	5
[1 0 0 0 0]	5.5	[0 0 0 1 0]	20	11.8	5
[1 0 0 0 0]	5.5	[0 0 1 0 0]	20	11.8	5
[1 0 0 0 0]	5.5	[0 1 0 0 0]	20	11.8	5
[1 0 0 0 0]	5.5	[1 0 0 0 0]	20	-9.1	-35
[1 0 0 0 0]	5.5	[0 0 0 1 1]	40	11.8	15
[1 0 0 0 0]	5.5	[0 0 1 1 0]	40	11.8	15
[1 0 0 0 0]	5.5	[0 1 1 0 0]	40	11.8	15
[1 0 0 0 0]	5.5	[1 1 0 0 0]	40	-6.1	-25
[1 0 0 0 0]	5.5	[0 0 1 1 1]	60	11.8	25
[1 0 0 0 0]	5.5	[0 1 1 1 0]	60	11.8	25
[1 0 0 0 0]	5.5	[1 1 1 0 0]	60	-4.4	-15
[1 0 0 0 0]	5.5	[0 1 1 1 1]	80	11.8	35
[1 0 0 0 0]	5.5	[1 1 1 1 0]	80	-3.2	-5
[1 0 0 0 0]	5.5	[1 1 1 1 1]	100	-2.3	5

Table 5.3: Rewards for each action, when the target is closer to the radar.

Observed Interference	Range (km)	Action	BW (MHz)	SINR (dB)	Reward
[1 0 0 0 0]	3.8	[0 0 0 0 1]	20	18.2	8
[1 0 0 0 0]	3.8	[0 0 0 1 0]	20	18.2	8
[1 0 0 0 0]	3.8	[0 0 1 0 0]	20	18.2	8
[1 0 0 0 0]	3.8	[0 1 0 0 0]	20	18.2	8
[1 0 0 0 0]	3.8	[1 0 0 0 0]	20	−2.7	−35
[1 0 0 0 0]	3.8	[0 0 0 1 1]	40	18.2	18
[1 0 0 0 0]	3.8	[0 0 1 1 0]	40	18.2	18
[1 0 0 0 0]	3.8	[0 1 1 0 0]	40	18.2	18
[1 0 0 0 0]	3.8	[1 1 0 0 0]	40	0.3	−25
[1 0 0 0 0]	3.8	[0 0 1 1 1]	60	18.2	28
[1 0 0 0 0]	3.8	[0 1 1 1 0]	60	18.2	28
[1 0 0 0 0]	3.8	[1 1 1 0 0]	60	2.0	22
[1 0 0 0 0]	3.8	[0 1 1 1 1]	80	18.2	38
[1 0 0 0 0]	3.8	[1 1 1 1 0]	80	3.2	32
[1 0 0 0 0]	3.8	[1 1 1 1 1]	100	4.2	42

5.2 Intermittent interference

The intermittent interferer model is similar to the constant interferer, except the interferer is no longer "on" for the entire duration of the training run. In these experiments, the radar was tested for 10% and 90% interference transmission probability. Note that the interference uses consistent frequency bands when transmitting during each training run. A higher percentage means the interferer is "on" for a greater amount of time. The probability of interference transmission is independent from one time instant to the next. This scenario is useful for modeling the performance of communication systems that occupy a specific frequency band, but whose transmissions can vary in duration. Results for the 10% case are shown in Figure 5.2. In this case, the radar has learned that interference is unlikely, and thus selects all bands for the entire simulation length. For the 90% case in Figure 5.3, the radar behaves similar to the constant interference case, where it avoids the interferer until the target is close, and then the radar selects all bands. In the 90% case, the radar learns that the interference is likely, and thus waits until the SINR is sufficiently high before maximizing its bandwidth. In both cases the radar learns that although it can't predict when interference will occur, it can learn the probability of interference.

In the 10% case, the rare penalty for negative SINR due to infrequent interference is tolerated in exchange for the benefit of using more bandwidth. When the interferer transmits more frequently, the penalty is more common and thus severe, thus the radar avoids the band that contains the interferer until the SINR is guaranteed to stay positive. In the 10% case, we can make the radar more reactive to the interferer by increasing the penalty for negative SINR. When the radar senses interference, the radar's selected action is to immediately avoid the interferer, and only return when the interferer stops transmitting. This could be problematic in a practical target tracking application, because the radar would spend more of its time switching between bands in an attempt to avoid the interferer. Note that in our model, a sense and avoid strategy (DSA) does not make sense, since the sensing interval is consistent with the interference duration. The interference can change over a sensing interval, and a DSA implementation could potentially use a band occupied by an interferer.

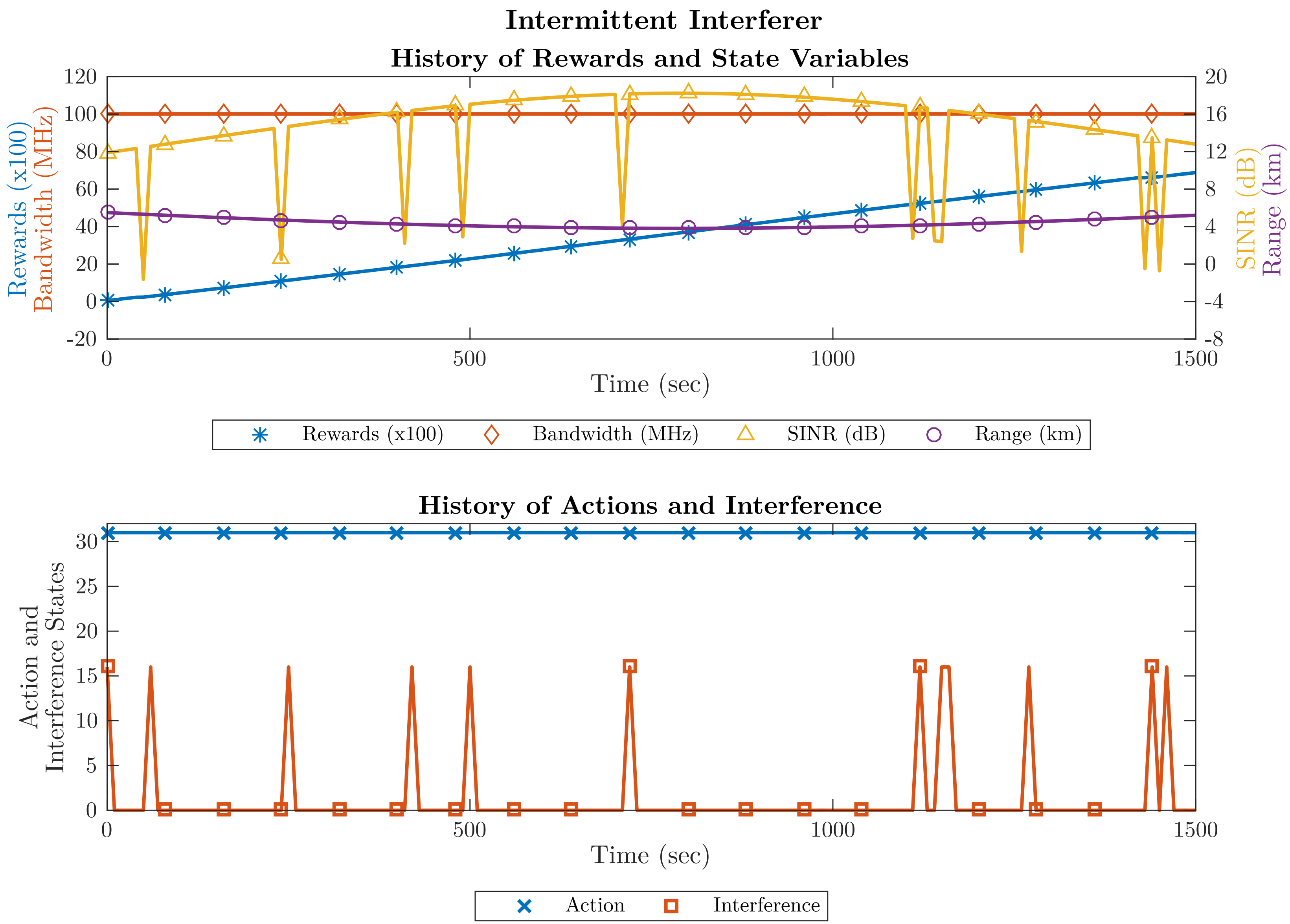

Figure 5.2: Results for 10% intermittent interferer.

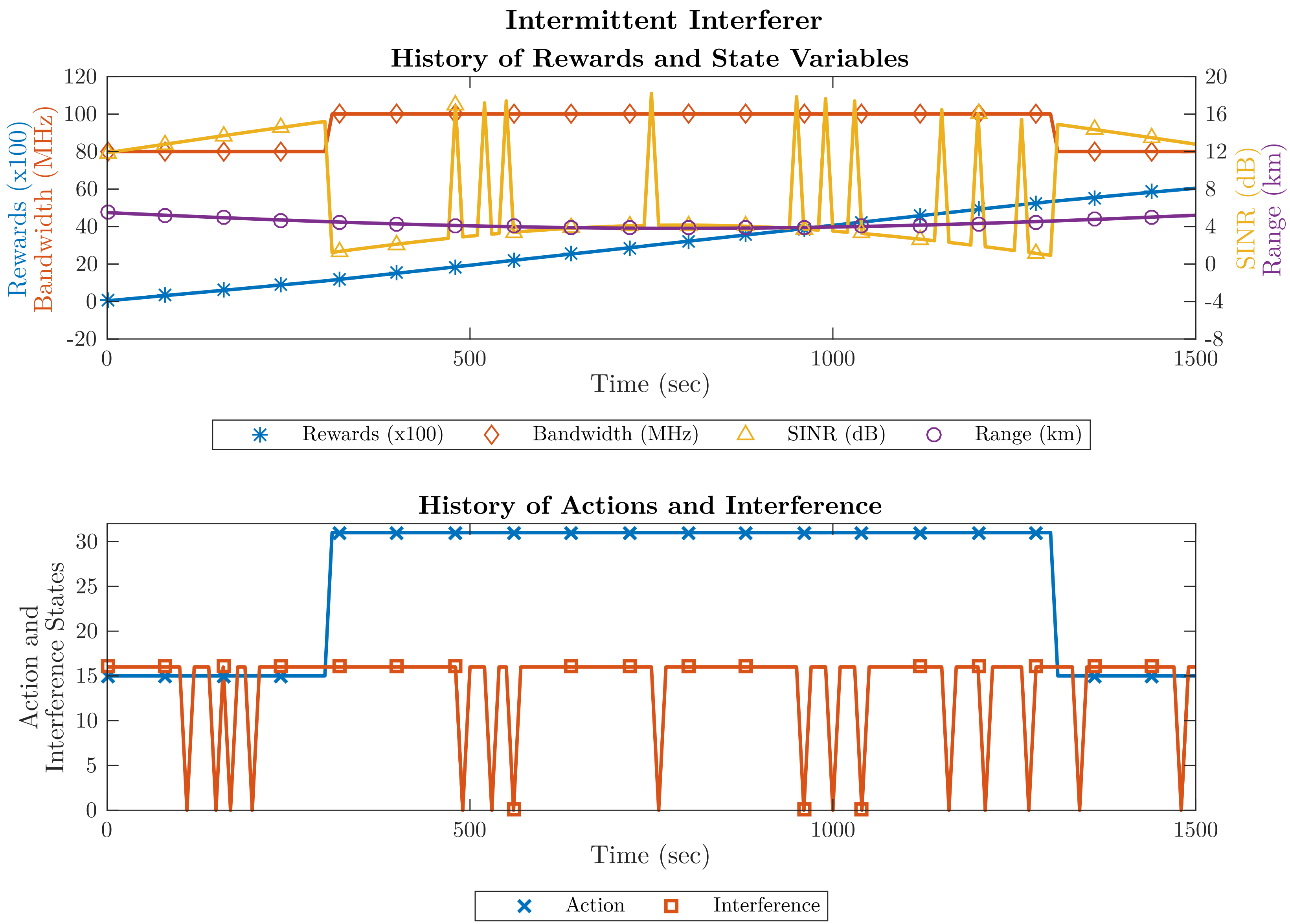

Figure 5.3: Results for 90% intermittent interferer.

Changing the reward structure, namely increasing or decreasing the penalty of the negative SINR with respect to the reward for using all of the bands, makes a difference in the radar's behavior. The current reward structure is set up such that the magnitude of the penalty for negative SINR is greater than the magnitude of the reward for bandwidth, except when the radar uses all bands. This setup discourages the radar from having a negative SINR in most cases, except when the radar uses all bands. Even if the radar uses all bands, if SINR would be negative, the radar will avoid this situation by using fewer bands. If the radar uses four bands or less and has a negative SINR, the value function will be negative and thus the optimal policy will be to use fewer bands, and avoid the interferer entirely. When the radar uses all of the available bands, the value function can be positive even when the SINR is negative. This does not necessarily mean using all bands is optimal, however. With this setup, the radar has some incentive (a small net reward of 40 - 35 = 5) to use all of the bands. If the magnitude of the penalty is greater than the magnitude of the reward for using all bands, (e.g. -50 instead of -35), there will no longer be any incentive for the radar to take risks by using all bands as the value functions will be negative when the SINR is negative, regardless of how many bands are used. As a result, the radar will be more "reactive"; *i.e.* whenever it senses interference, it will immediately adjust its band usage to avoid the interferer, wait there for as long as the interferer is there, and then move back to the bands it was using before the interference appeared.

5.3 Triangular frequency sweep

The triangular frequency sweep interferer occupies one band at a time and moves up and down the available bands, creating a triangular wave pattern when viewed on a waterfall plot as shown in Figure 5.4. This case (along with others to be considered) models the radar's performance in the presence of a deterministic frequency hopping communications system.

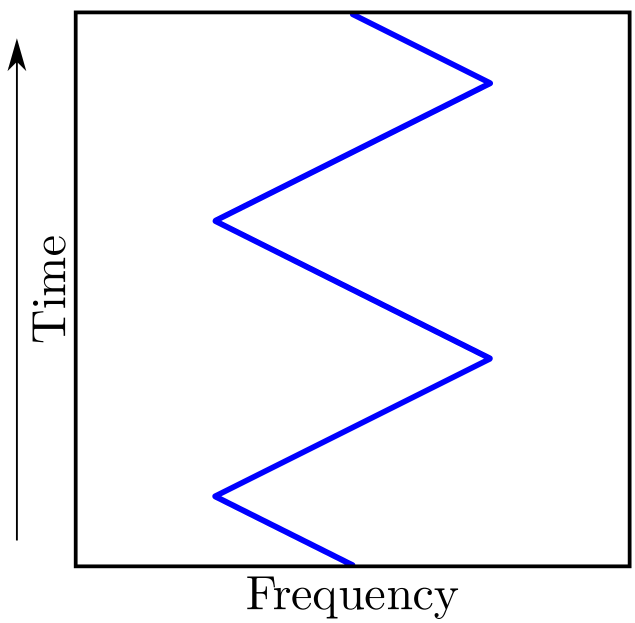

Figure 5.4: Waterfall plot of triangular frequency sweep interferer.

The results shown in Figure 5.5 demonstrate that the SINR fluctuates greatly as the interferer moves around in frequency. When the interferer occupies any of the middle three bands, the radar is not able to predict where the interferer will go next since the radar does not know whether the frequency is increasing or decreasing based on only the current band. Only when the the interferer is at the edge of the available bands, does the radar know where the interferer will go next. As a result, the radar's policy is to maximize bandwidth, even if there is a risk of collision with an interferer, because avoiding interference would mean reducing the bandwidth too much (only a one band waveform could avoid interference entirely), thus resulting in fewer rewards. This behavior depends on 1. the specific range of the target (i.e. its SINR), and 2. the penalties/rewards structure for negative SINR/bandwidth, respectively. The behavior of the radar given the observed interference is outlined in Table 5.4. When the interferer is at the band edges ($\boldsymbol{\theta} = [00001]$ or $[10000]$), the radar knows with certainty the future interference state will be [00010] or [01000], respectively and uses the three bands where the interferer will not go [11100] or [00111], respectively. When the interferer occupies any of the middle three bands ($\boldsymbol{\theta} = [00100]$, for example), the radar has learned there is an equal probability (50%) of the interferer's future state being either [01000] or [00010]. Due to the set up of the reward structure, it is advantageous for the radar to use all of bands even if the SINR will be negative because the reward is greater than if the radar attempted to avoid the interferer by using four bands (+5 from Equation A.11 vs. +1 from Equation A.13).

To improve the performance we need to increase the number of states to include the previous two interference states. Specifically, we modify the model to include the current interference state at time t and the previous interference state at $t-1$. The number of states becomes $N_S = \rho\nu 2^N \cdot 2^N = \rho\nu 2^{2N}$, an increase by a factor of 2^N. When memory is employed in this way (see Figure 5.6 and Table 5.5), the radar is able to predict where the interferer will go, and therefore there are no drops in the SINR. The cost for using memory is training time and complexity. Specifically, training becomes somewhat longer because more training runs are needed to cover the increase in the number of states. When memory is used, the radar knows what the future interference state will be, given the current observed and previous states. Table 5.5 demonstrates the radar has learned the interference behavior because each action optimizes the bandwidth it can use while also keeping the SINR positive. This coincides with the result in Figure 5.6, as the SINR never drops below 0 dB.

For more detail about how some of the value functions for the triangle frequency sweep case are computed, the reader is directed to Table A.3 for when memory is not used, and Table A.4 for when memory is employed. The value functions are computed by looking at the reward obtained when the radar takes an action and transitions from the current state to the future state(s).

Table 5.4: Interference states and actions for triangle sweep interferer, without memory.

Observed Interference	Action	Future Interference
00001	11100	00010
00010	11110	00100
00100	11111	01000
01000	01111	10000
10000	00111	01000
01000	01111	00100
00100	11111	00010
00010	11110	00001
00001	11100	00010

Table 5.5: Interference states and actions for triangle sweep interferer, with memory.

Observed Interference	Action	Future Interference
00001	11100	00010
00010	11000	00100
00100	00111	01000
01000	01111	10000
10000	00111	01000
01000	00011	00100
00100	11100	00010
00010	11110	00001
00001	11100	00010

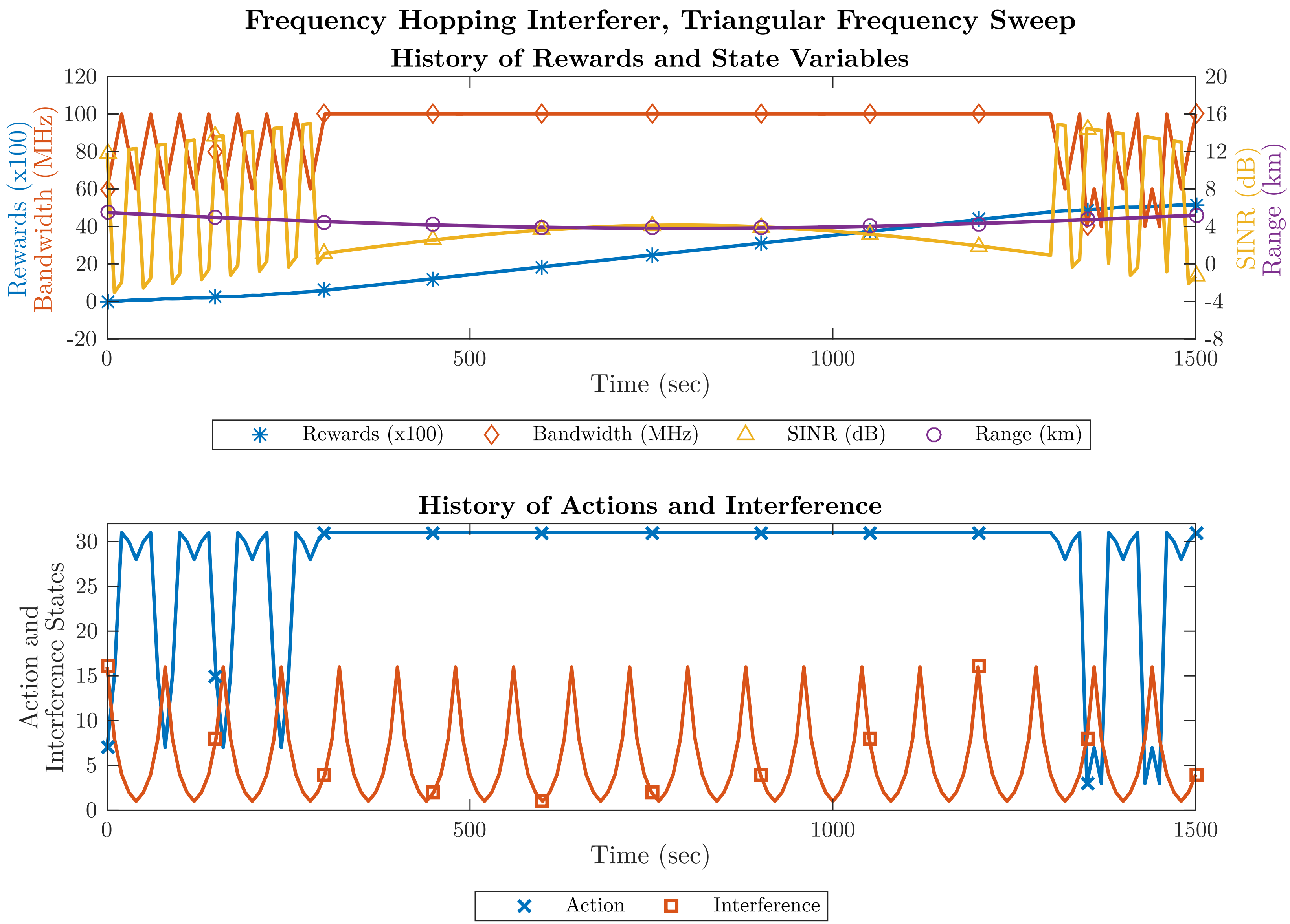

Figure 5.5: Results for triangle sweep interferer, without memory.

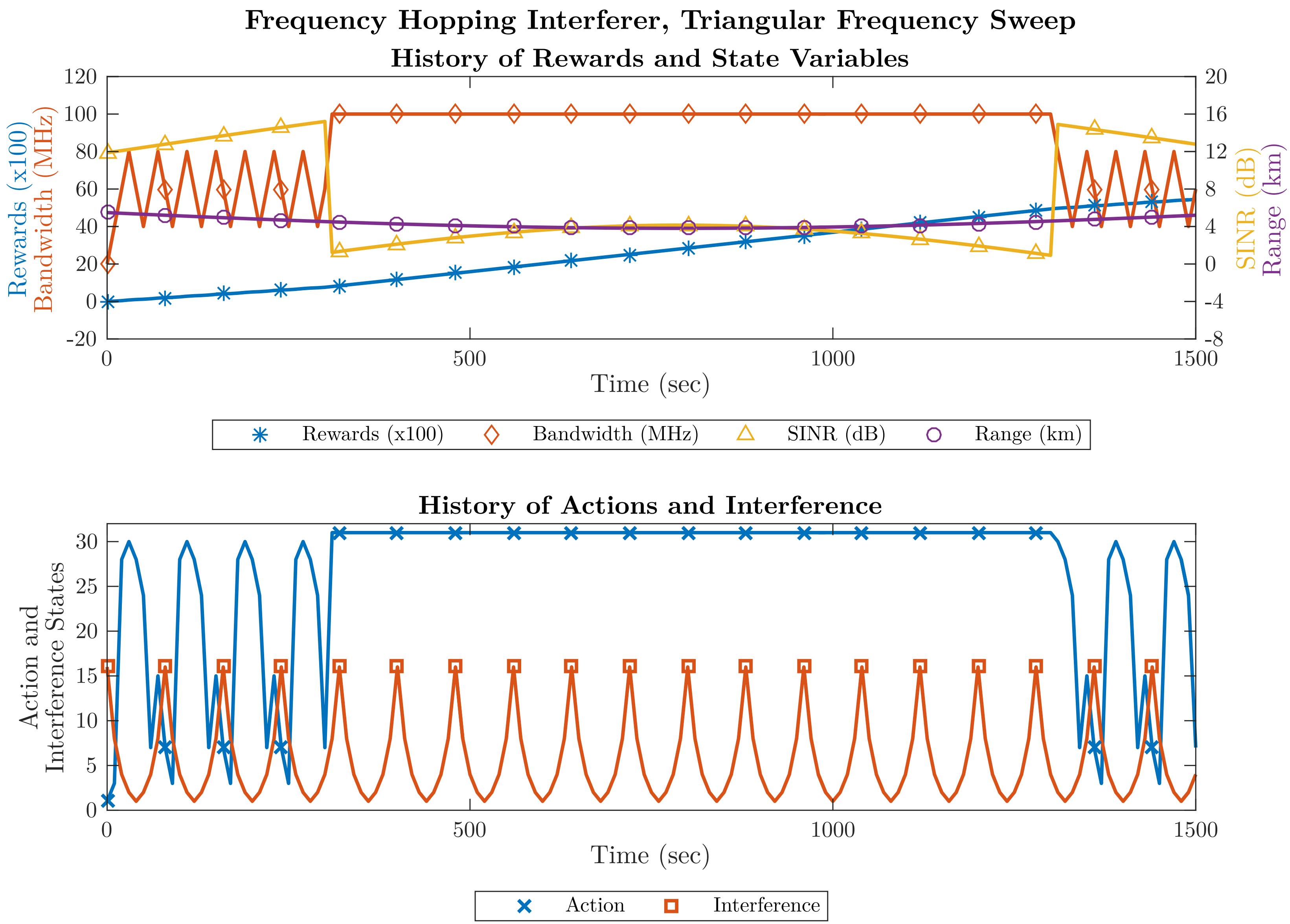

Figure 5.6: Results for triangle sweep interferer, with memory.

5.4 Sawtooth frequency sweep

The sawtooth frequency sweep interferer occupies one band at a time, and moves in one direction, wrapping around when the band edge is reached. The sawtooth pattern is illustrated in Figure 5.7. Like the triangular frequency sweep, this interferer is useful for evaluating the radar's performance in the presence of deterministic frequency hopping communications systems.

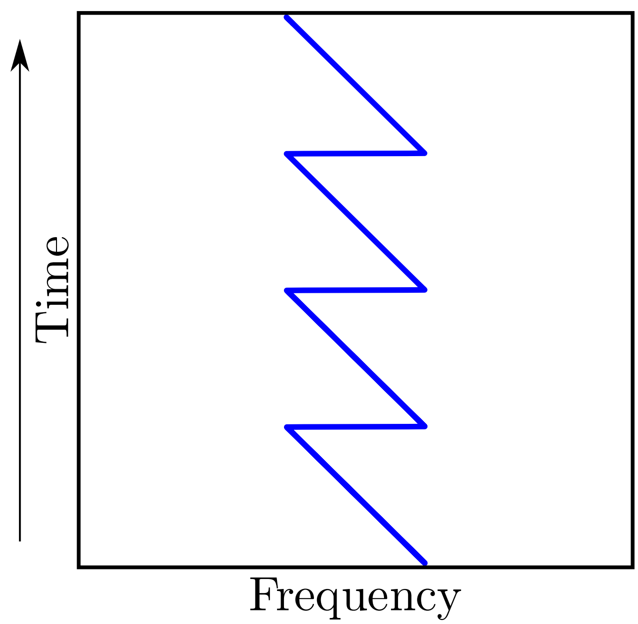

Figure 5.7: Waterfall plot of sawtooth frequency sweep interferer.

The results in Figure 5.8 demonstrate the radar is able to predict where the interferer is going to go, thus avoiding any drops in SINR. Unlike the triangle frequency sweep case, the transition probability from one interference state to the next is $\mathbb{P}(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) = 1.0$, thus the radar knows which bands the interferer will use, and adjusts accordingly. When the target is close enough and SINR is sufficiently high, the radar then selects all bands, as the radar can accept the lower (but still positive) SINR, to get more reward from bandwidth.

Table 5.6 shows the actions taken by radar for each observed interference state. Unlike the triangle frequency sweep, the transition probability in any state is 1, instead of 0.5. The radar is thus able to learn and predict the behavior of the interferer, and it chooses to use as many bands as possible while also avoiding the interferer. For example, if the current state is $\boldsymbol{\theta} = [10000]$, the future interference state will be $\boldsymbol{\theta} = [01000]$ with a transition probability of 1. The action the radar selects for the future state is $a = [00111]$, which is optimal because it uses the most contiguous amount of bands, while avoiding the interferer, which keeps the SINR positive.

Table 5.6: Interference states and actions for sawtooth sweep interferer.

Observed Interference	Action	Future Interference
10000	00111	01000
01000	00011	00100
00100	11100	00010
00010	11110	00001
00001	01111	00010

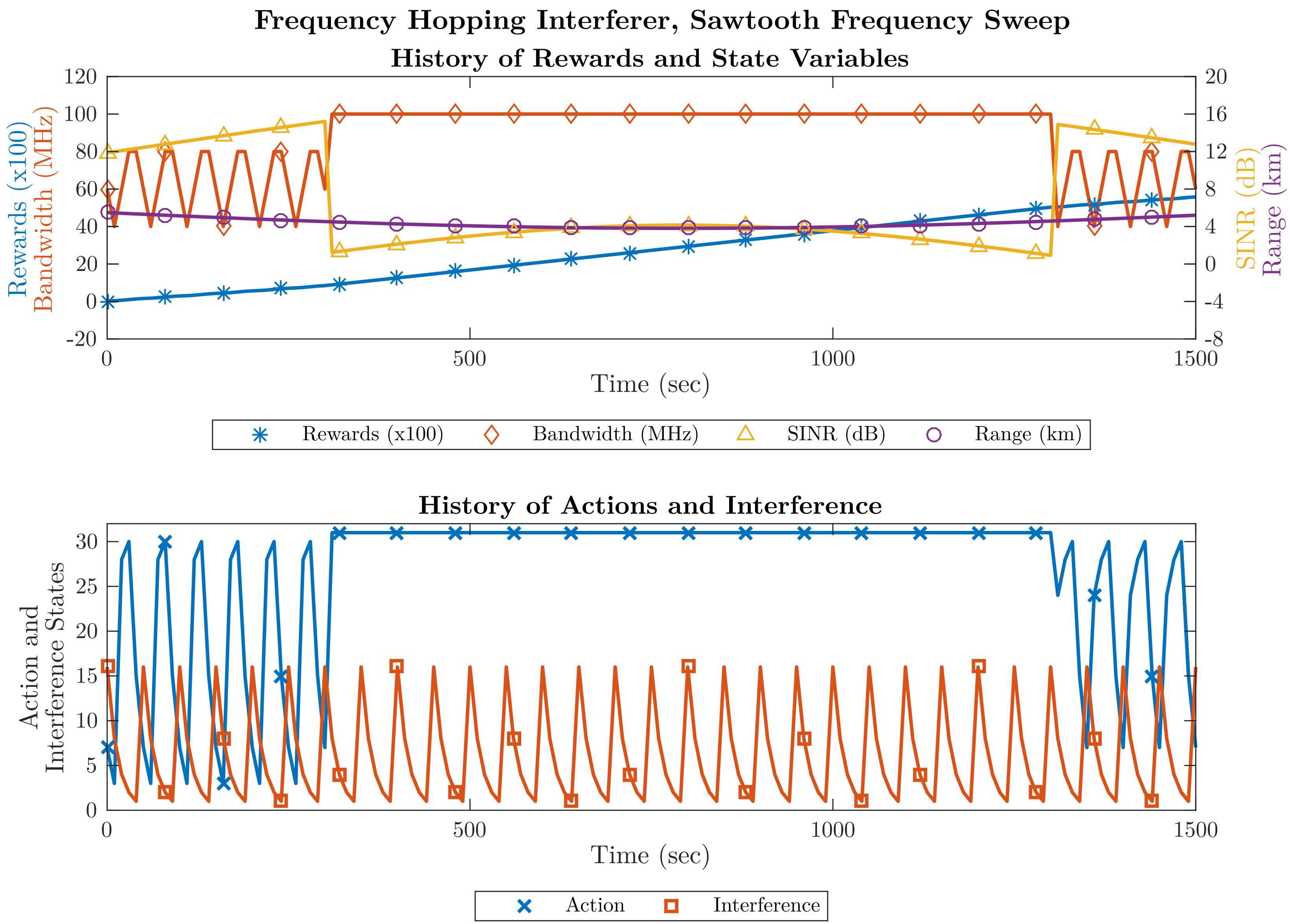

Figure 5.8: Results for sawtooth frequency sweep interferer.

5.5 Length-5 pseudorandom frequency hop

The length-5 frequency hopper occupies one band at a time, but unlike the triangle and sawtooth frequency sweep, doesn't always move to neighboring bands. The hopping pattern used in this case $\{3, 1, 2, 4, 5, \ldots\}$. When the last band in the sequence is reached, the interferer goes back to the first band, and the sequence repeats. This case is useful for modeling the performance of the radar in the presence of short pseudorandom frequency hopping communication systems. As the results in Figure 5.9 show, the radar has learned the optimal behavior, as it is able to predict and avoid the interferer's movements, and use all of the bands only when the target is close to the radar.

Table 5.7 lists the action taken for each observed interference state. Since the transition probability from one state to the next is 1 (unlike triangle sweep, without memory), the radar learns with certainty what the next interference state is going to be. As a result, the action selected for the future state, given the current state, uses as many bands as possible while avoiding the interferer and preventing the SINR from becoming negative.

Table 5.7: Interference states and actions for length-5 frequency hopping interferer.

Observed Interference	Action	Future Interference
00100	01111	10000
10000	00111	01000
01000	11100	00010
00010	11110	00001
00001	11000	00100

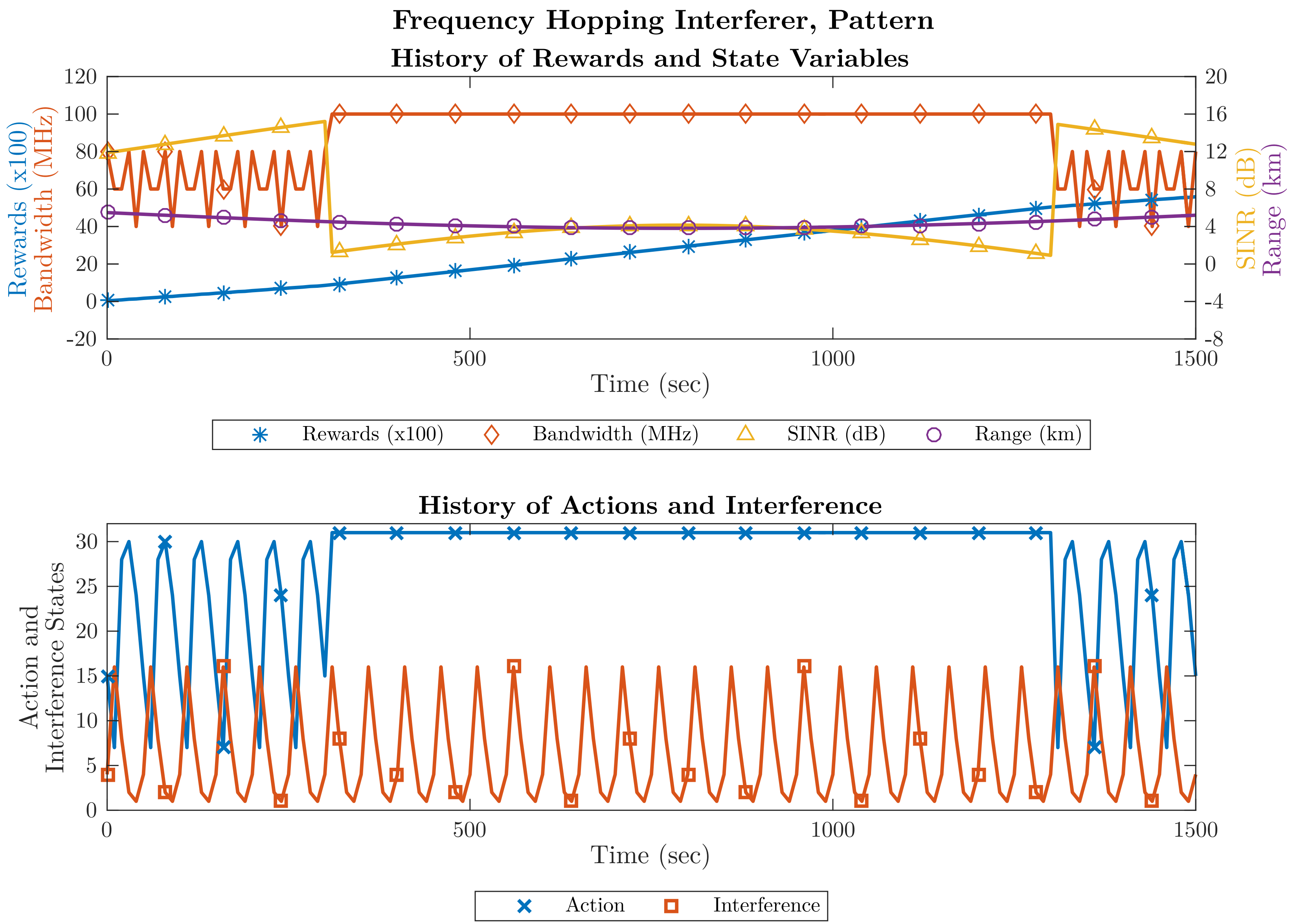

Figure 5.9: Results for length-5 frequency hopping interferer.

5.6 Length-10 pseudorandom frequency hop

The length-10 frequency hopper occupies one band at a time, and like the length-5 frequency hop, doesn't always move to neighboring bands. The hopping pattern used in this case is $\{3, 1, 2, 4, 5, 2, 3, 4, 1, 5, \ldots\}$. When the last band in the sequence is reached, the interferer goes back to the first band, and the sequence repeats. Again, this case is useful for modeling the performance of the radar in the presence of frequency hopping spread spectrum communications systems.

Unlike the length-5 frequency hopper, the transition probability (as seen from the radar's perspective) from one interference state to the next is going to be $\mathbb{P}(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) = 0.5$. As the results in Figure 5.10 demonstrate, the radar is unable to predict the future interference state, similar to the triangle frequency sweep interferer. For example, if the current interference state is $\boldsymbol{\theta}_t = [1\,0\,0\,0\,0]$ the future interference state could be either $\boldsymbol{\theta}_{t+1} = [0\,1\,0\,0\,0]$ or $\boldsymbol{\theta}_{t+1} = [0\,0\,0\,0\,1]$, both with equal transition probabilities of 0.5. The action selected in this instance is $\boldsymbol{a}_t = [1\,1\,1\,1\,0]$, which is optimal given the scenario because the radar attempts to maximize the reward it can get from bandwidth, despite the 50% probability of using the same bands as the interferer and having a negative SINR. The radar occasionally is able to successfully avoid the interferer, but this only occurs when the occupied bands of the future states are next to each other. For example, when $\boldsymbol{\theta}_t = [0\,0\,0\,0\,1]$ and could transition to either $\boldsymbol{\theta}_{t+1} = [0\,1\,0\,0\,0]$ or $\boldsymbol{\theta}_{t+1} = [0\,0\,1\,0\,0]$, the radar selects $\boldsymbol{a}_t = [0\,0\,0\,1\,1]$, because it knows either bands 2 or 3 will be occupied, and bands 4 or 5 provide an opportunity to maximize bandwidth and a guarantee of not colliding with the interferer.

Because the radar is unable to predict what the interferer state will be, its performance is suboptimal and suffers drops in SINR. Like the triangle frequency sweep case, we can also utilize memory, such that each state contains information of the interference state on the current and previous time steps. When memory is employed, the transition probabilities resolve to $\mathbb{P}(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = 1$, which means the radar knows with certainty which interference state is next. The results, in Figure 5.11, demonstrate the optimal performance of the radar when memory is used.

Table 5.8 shows the actions taken by the radar given the current interference state when memory is not used. Similar to the triangle frequency sweep case, each state has two possible future states, each with transition probabilities of 50%. Since the radar does not know which future state is more likely, the general behavior of the radar is to use as many bands as possible to maximize the reward from bandwidth, even if there is a 50% risk of negative SINR. As a result, there are times at which the SINR goes negative because the radar uses one of the bands occupied by the communications system.

Table 5.9 shows the actions the radar selects when memory is employed in the length-10 psedorandom hop. Given the current interference state, and the previous state, the radar knows with certainty what the future state will be. Therefore, each action selects as many bands as possible, while also avoiding the interferer in the future state. The actions in the table coincide with the results in Figure 5.11.

Table 5.8: Interference states and actions for length-10 frequency hopping interferer, without memory.

Observed Interference	Action	Future Interference
00100	01100	10000
10000	11110	01000
01000	11111	00010
00010	01110	00001
00001	00011	01000
01000	11111	00100
00100	01100	00010
00010	01110	10000
10000	11110	00001
00001	00011	00100

Table 5.9: Interference states and actions for length-10 frequency hopping interferer, with memory.

Observed Interference	Action	Future Interference
00100	01111	10000
10000	00111	01000
01000	11100	00010
00010	11110	00001
00001	00111	01000
01000	11000	00100
00100	11100	00010
00010	01111	10000
10000	11110	00001
00001	00011	00100

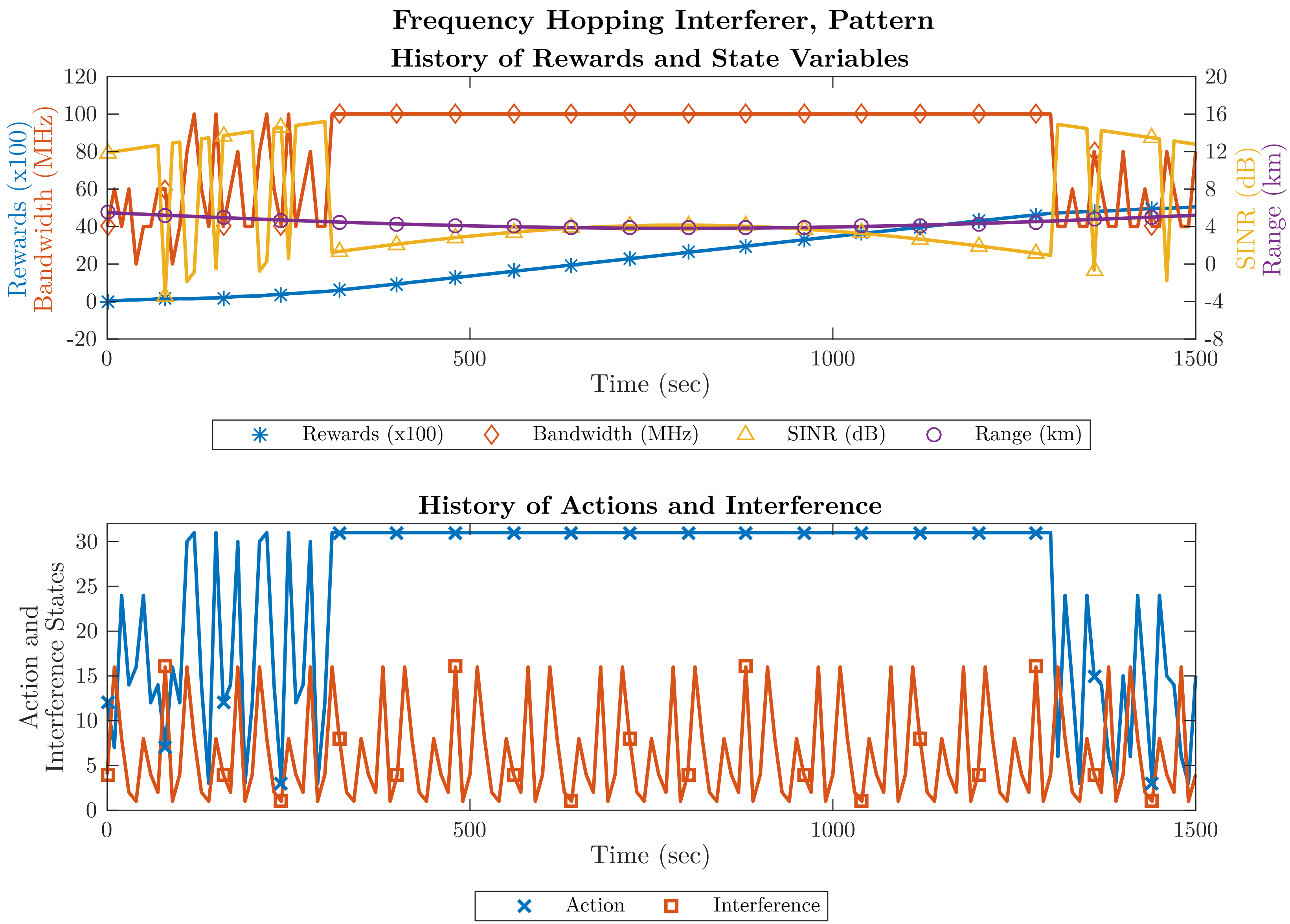

Figure 5.10: Results for length-10 frequency hopper, without memory.

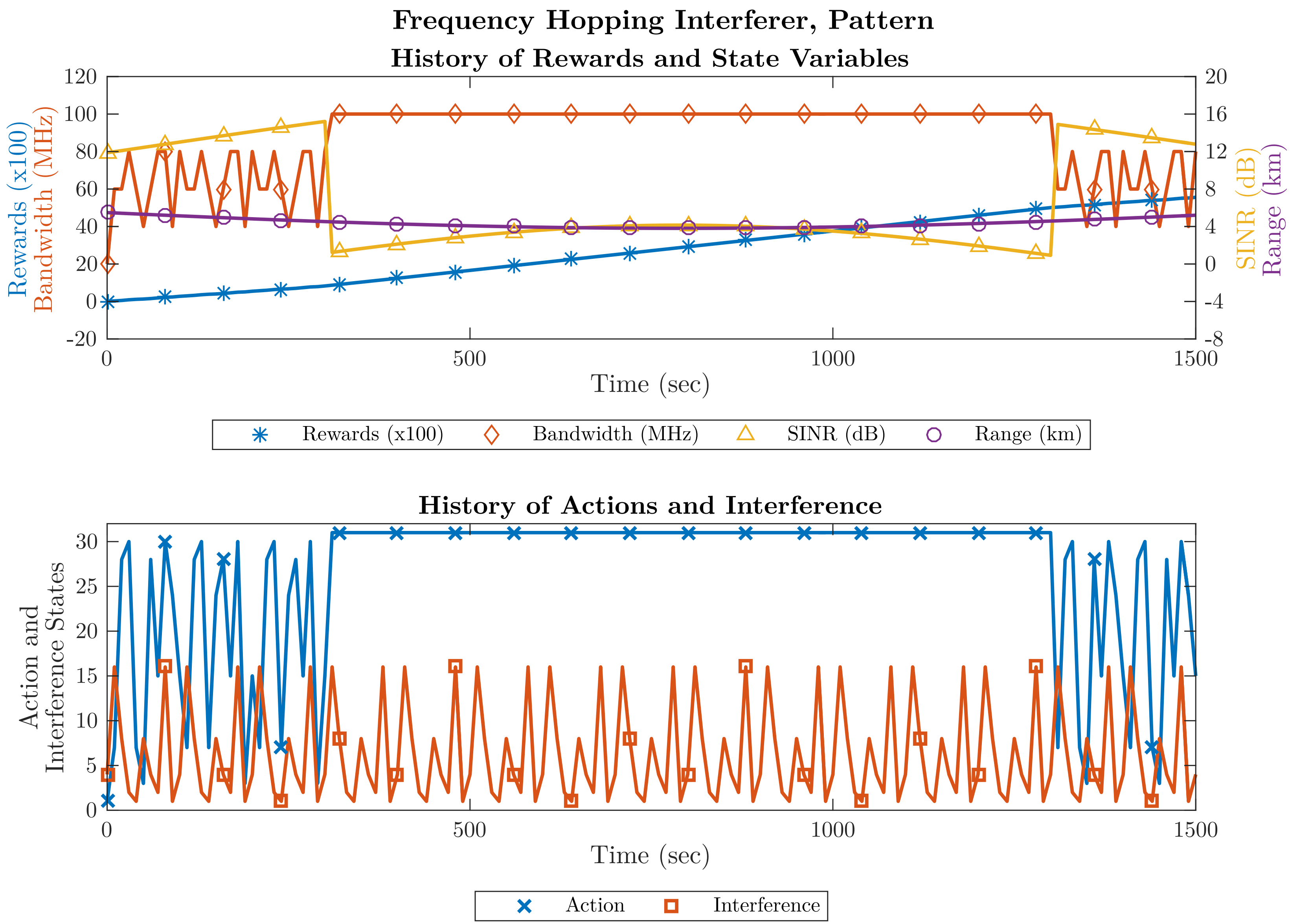

Figure 5.11: Results for length-10 frequency hopper, with memory.

5.7 Pseudorandom frequency hop

The pseudorandom frequency hop also occupies one band at a time, but unlike the previous two cases, it is a very long pseudorandom hop sequence. The transition probabilities from the current interference state to the next becomes uniformly distributed, and is the inverse of the number of bands, or in other words, $\mathbb{P}(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) = 1/N = 1/5$. As the results in Figure 5.12 demonstrate, the radar is unable to predict which bands the interferer will use, and therefore uses all of the bands all of the time, attempting to maximize reward from bandwidth, even if the SINR is low or negative. Like the other results, the radar does use all bands when the target is closer.

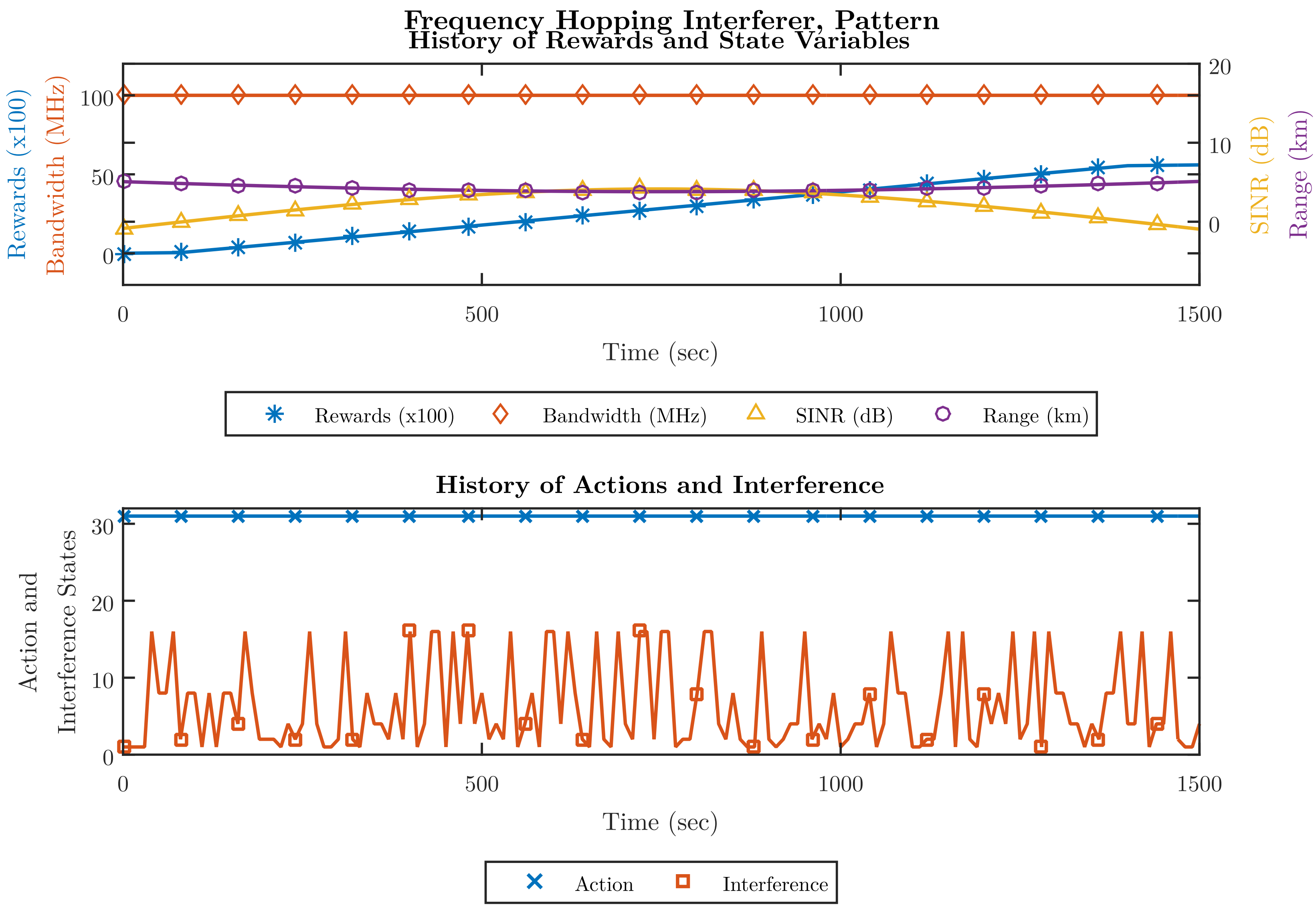

Figure 5.12: Results for pseudorandom frequency hop interferer.

5.8 Position-dependent interferer

The direction-dependent interference scenario, unlike the previous cases, removes the assumption that the interferer's power as sensed by the radar is direction independent. As Figure 5.13 illustrates, in this case the interference is localized. The interference affects only the position cells inside the red-dashed rectangle. When the radar's beam is tracking the target and the target is in these position states, the radar will also sense interference. When the radar is focused on the unaffected cells, it will not sense interference. We tested the radar against three cases in which the interferer is: 1. Constant, 2. Intermittent with high transmission probability (90%), and 3. Intermittent with low transmission probability (10%). Figure 5.14 shows the results for the constant interferer. When the target is in the unaffected cells, the radar learns to use all bands because there is no interference sensed, and thus all bands are available. Immediately before the target enters the regime with the affected position cells, the radar switches and avoids the band where the interferer resides. Before the target leaves the cells affected with interference, the radar selects to use all the bands again because it anticipates the target leaving the affected area. Because the position cells with interference is constant on each training run, the radar learns which cells (and what bands) will have interference, and is thus able to predict which cells will have interference and avoid those bands accordingly. By avoiding the interferer, the radar does not incur momentary drops in SINR. Note the decreasing SINR shown in the plot is only due to the target moving away from the radar and not the interferer. This behavior is optimal because the radar takes advantage of the bands being unoccupied by using all of them, and then avoiding the interferer by using one less band to maintain a positive SINR.

With the high probability of transmission intermittent interferer (Figure 5.15), the radar performs similar to the constant case, choosing the avoid the interferer when the target is in the affected position cells, and using all bands otherwise. Again, the radar avoids the band used by the interferer, and avoids momentary drops in SINR.

With the low probability of transmission intermittent interferer (Figure 5.16), the probability of transmission is low enough that the risk of having negative SINR is also low. It is therefore optimal for the radar to use all of the bands for the entire track, because it will receive more reward than if it attempted to avoid the (relatively) unlikely chance the interferer may transmit and cause the SINR to drop. Note that due to the intermittent nature of the interference, the radar cannot predict when interference will occur.

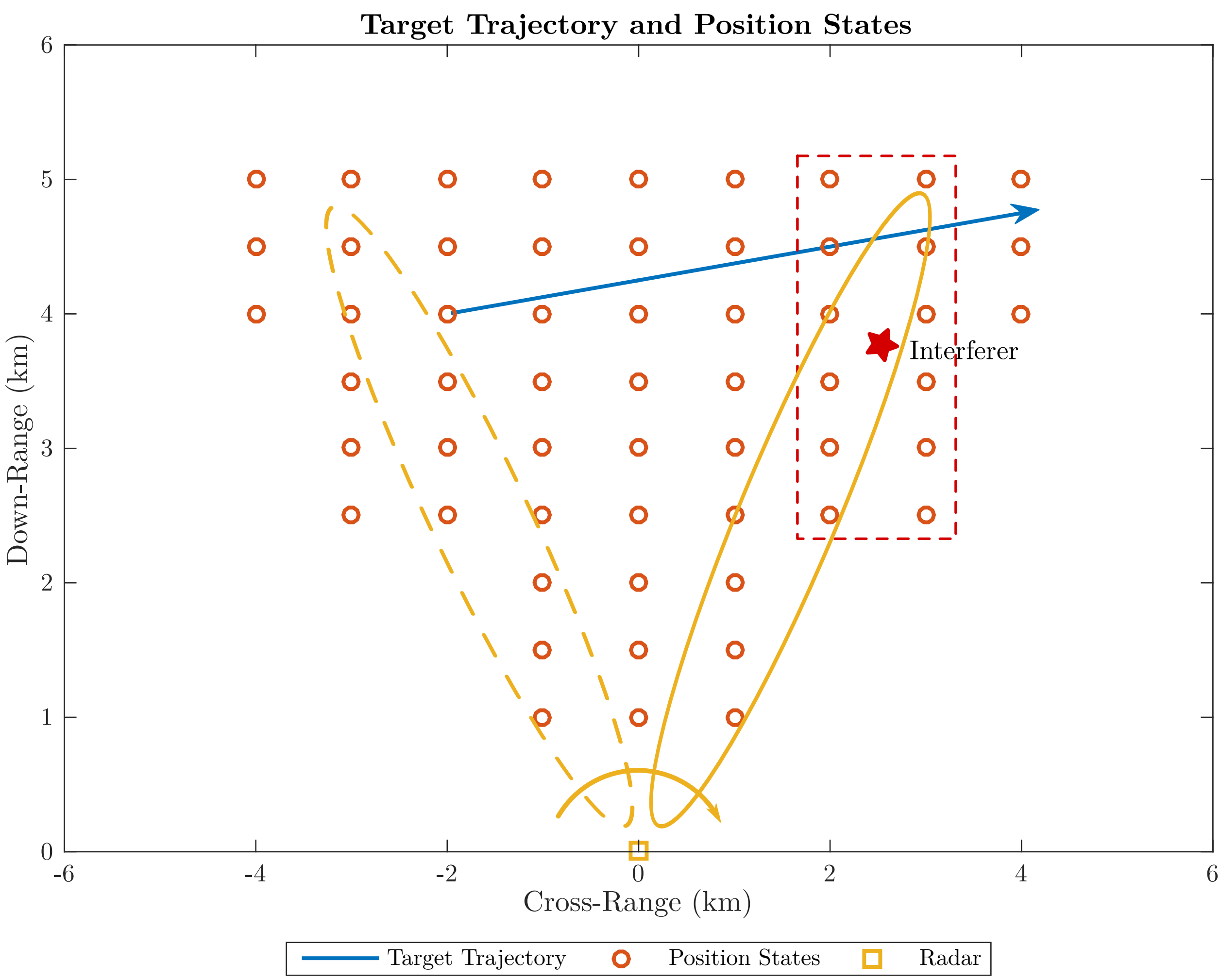

Figure 5.13: Trajectory of target with direction-dependent interferer.

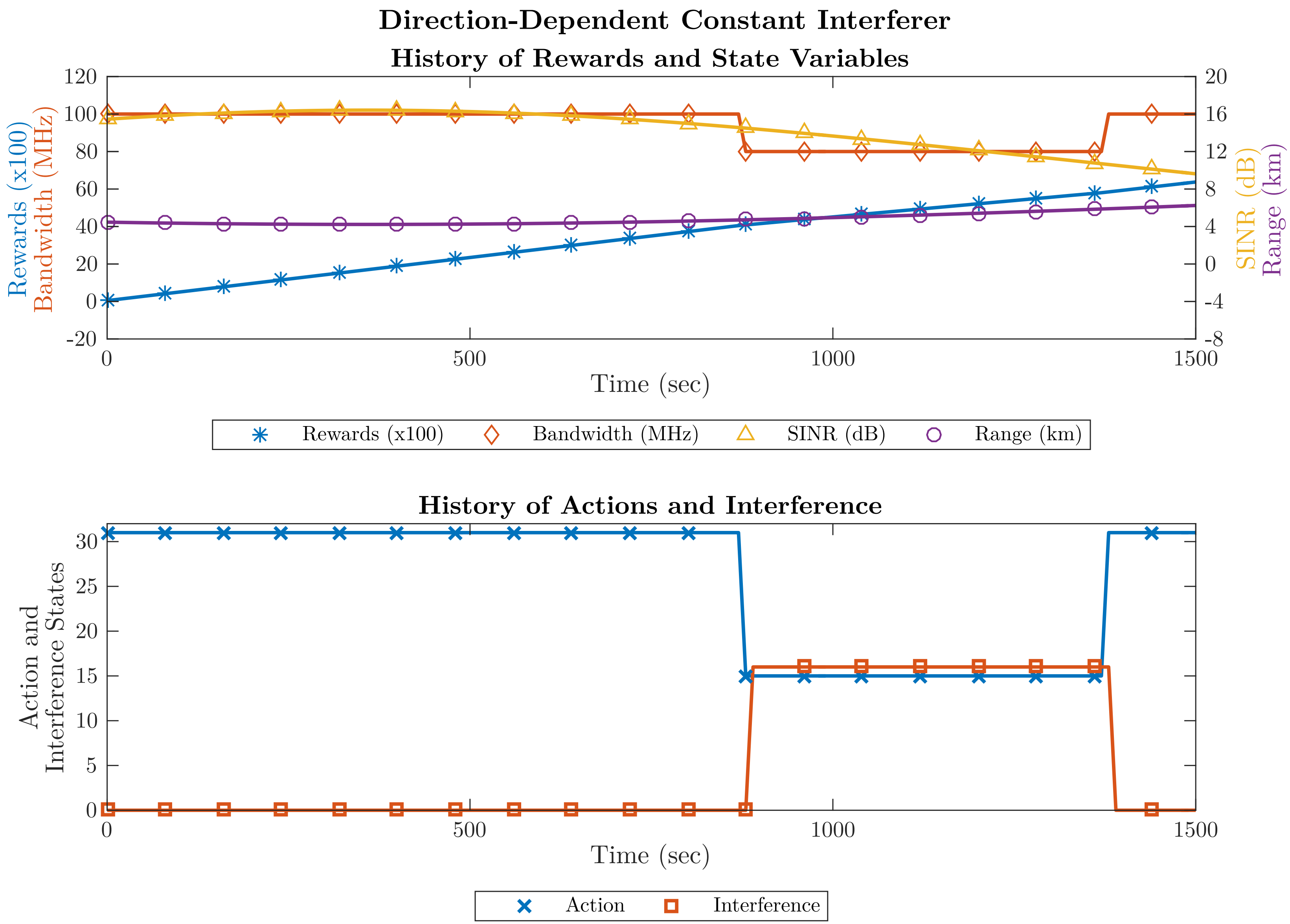

Figure 5.14: Results for direction-dependent constant interferer.

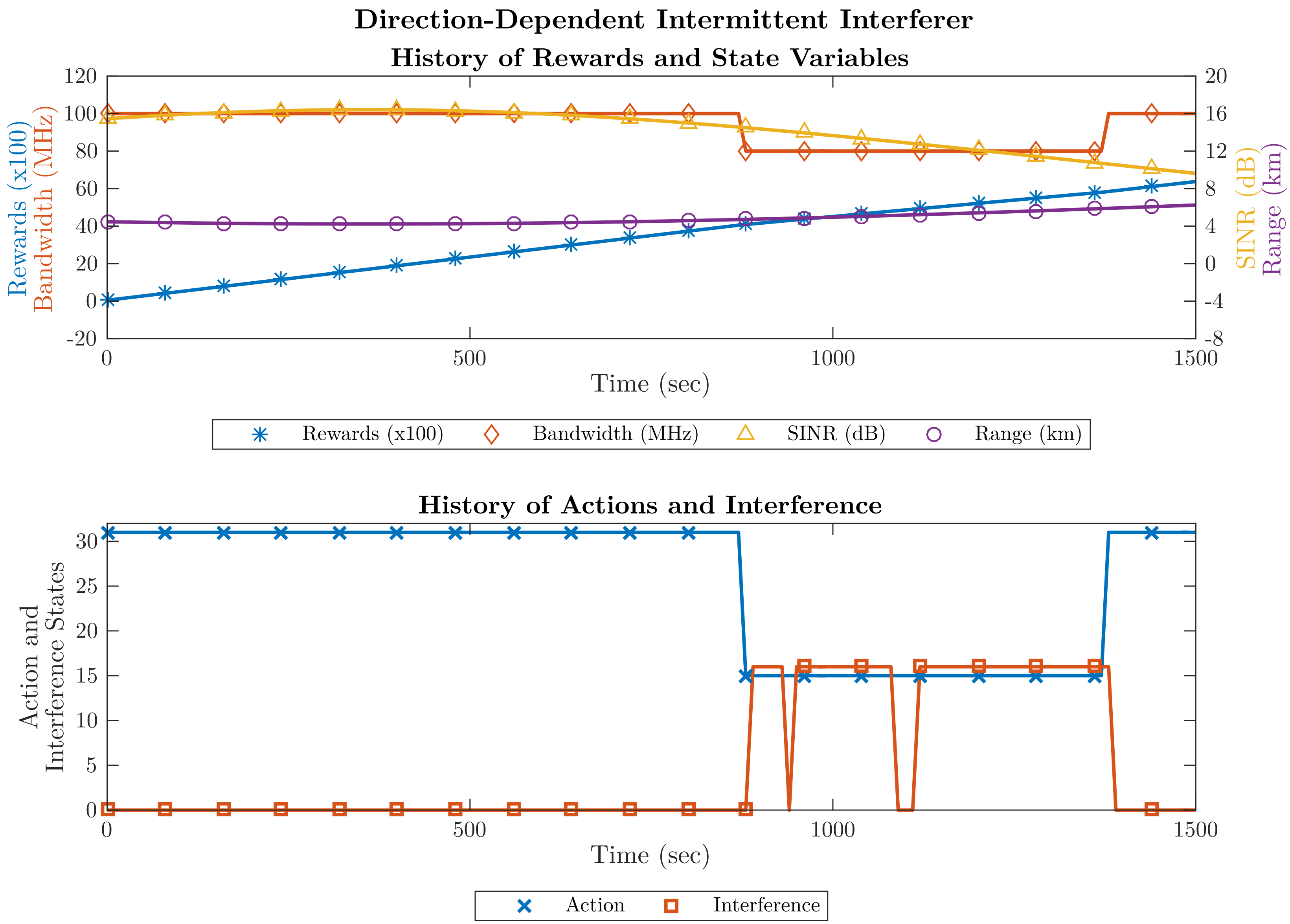

Figure 5.15: Results for direction-dependent intermittent interferer, with 90% transmission probability.

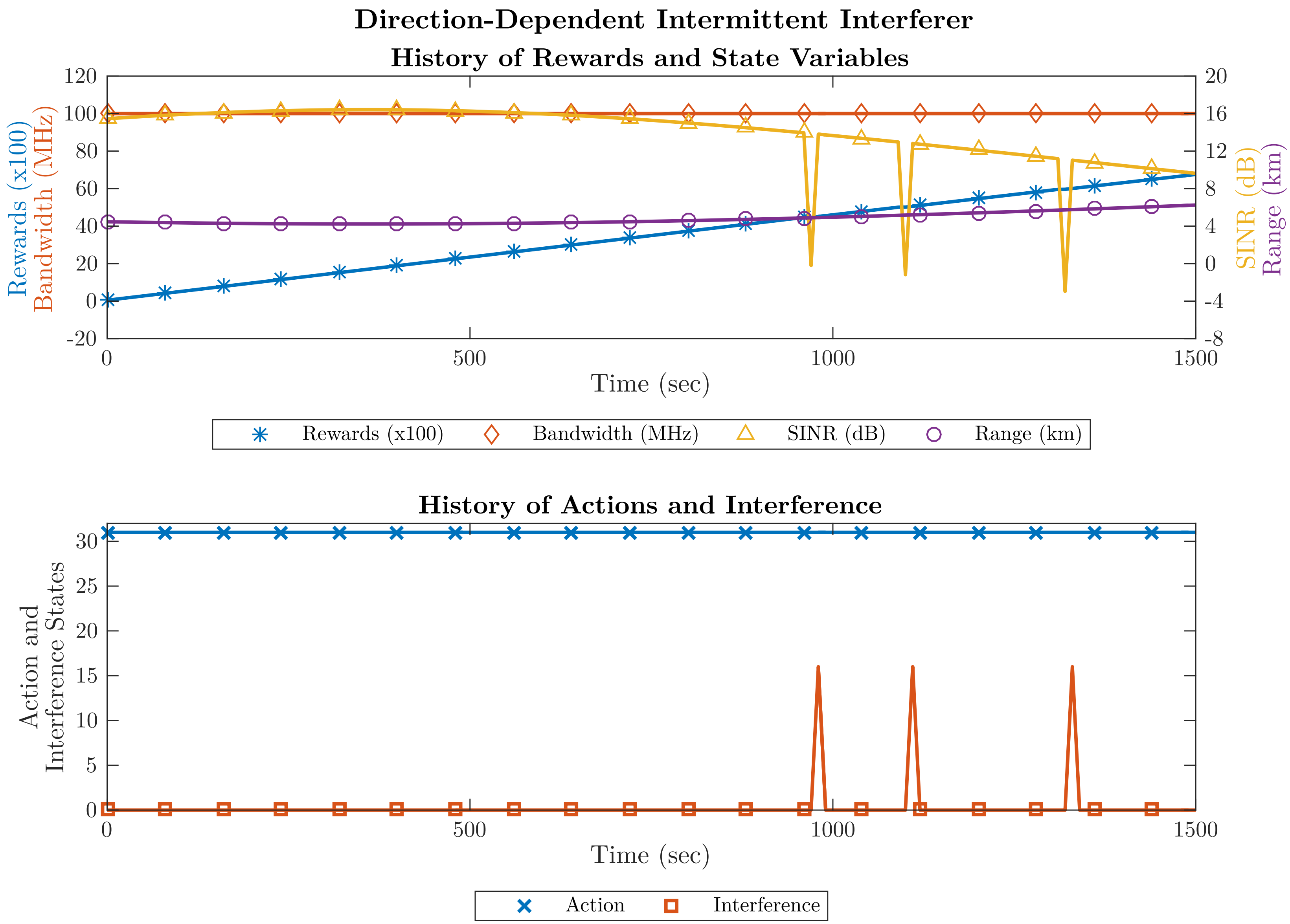

Figure 5.16: Results for direction-dependent intermittent interferer, with 10% transmission probability.

5.9 Comparison With Other Techniques

The following results compare the MDP model against another technique, dynamic spectrum access (DSA). A DSA system senses the spectrum and selects the bands available at that time. While DSA is simpler, it does not have the predictive ability of a radar modeled with an MDP and trained with reinforcement learning. Figure 5.17 compares the MDP and DSA models for the high intermittent (90%) case. The performance of the MDP model is indicated by the dotted lines and the DSA model is indicated by the dashed lines. The two approaches are compared by the reward accumulated at the end of the simulation. A DSA system is reactive to the interference, using bands only when they're unoccupied, but results in drops in SINR when the interferer transmits again. And since the DSA system is not learning from its environment, it does not use more bandwidth when the target is closer. With the same reward structure used in all prior results, the MDP model accumulates more reward than the DSA approach (6041 versus 5507).

Figure 5.18 shows the MDP and DSA approaches for the triangle sweep case. Comparing the accumulated reward, we see the MDP model outperforms the DSA system (5447 versus 1729). The difference in reward is due to the reinforcement learning, which enables the radar to (1) predict which bands the interferer will use in advance, and (2) learn at which target range can the radar trade SINR off for more bandwidth, thereby attaining better range resolution while maintaining positive SINR (both which maximize reward). Results for other cases (constant, low intermittent, sawtooth, etc.) also indicate the MDP model has a higher accumulated reward compared to the DSA technique, thereby demonstrating its superior performance.

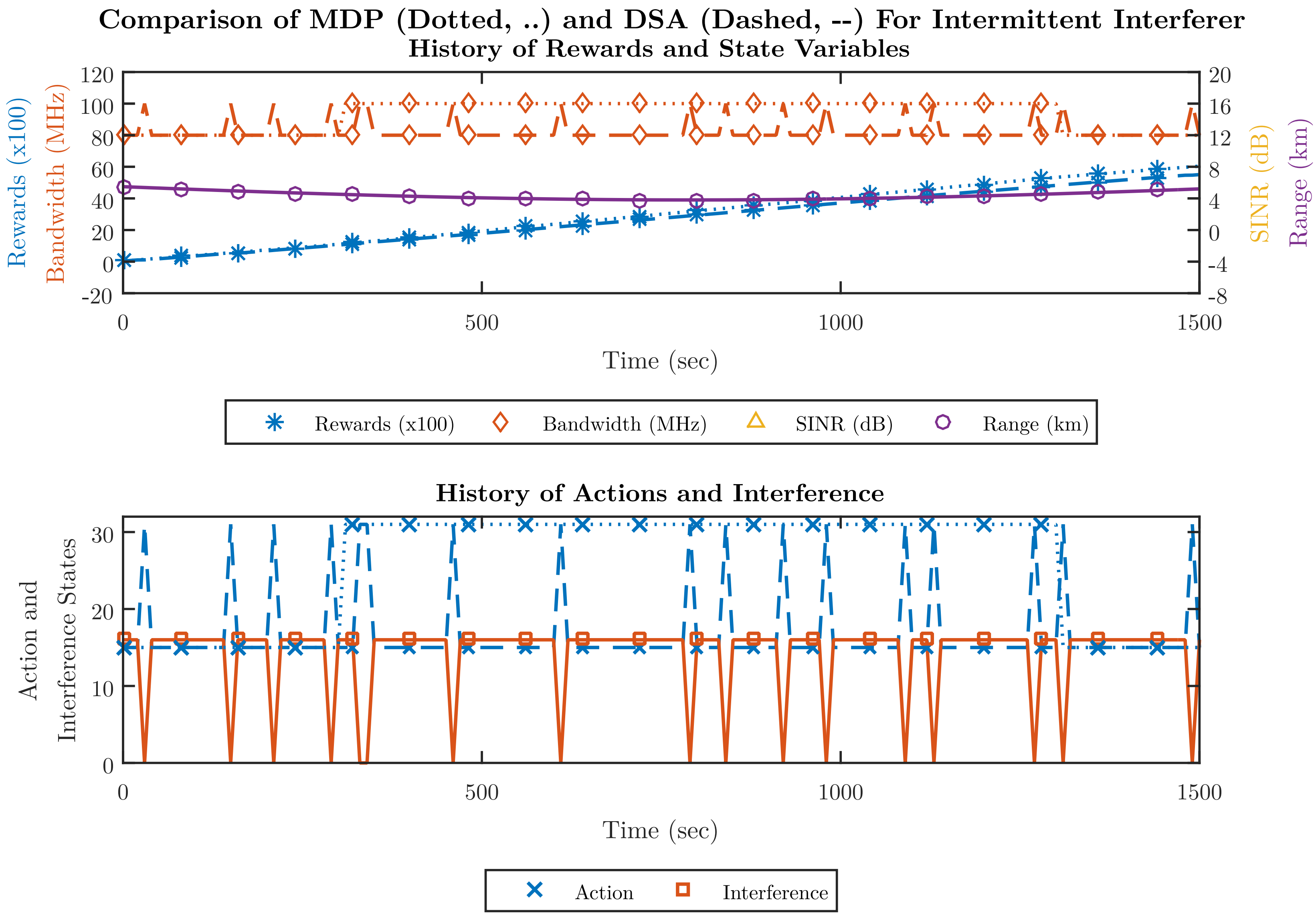

Figure 5.17: Results for comparing MDP and DSA for high intermittent case.

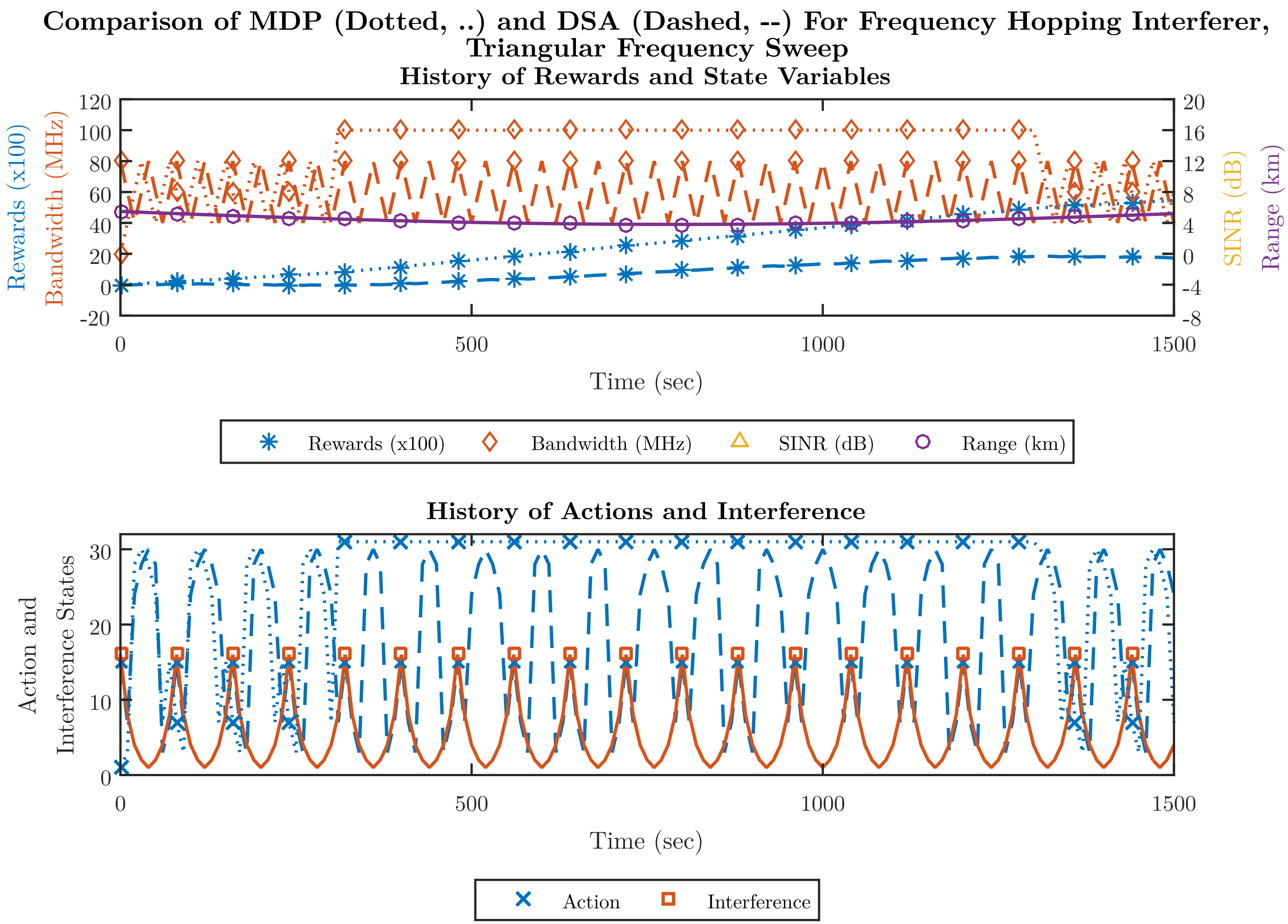

Figure 5.18: Results for comparing MDP and DSA for triangle frequency sweep case.

Chapter 6

Conclusion

This work has demonstrated the applicability of Markov decision processes and reinforcement learning to the radar-coexistence problem. The results demonstrate a radar is able to learn the interference behavior of a communication system and adjust its band usage to maintain the target track with a positive SINR. Building off of prior work in online optimization to adjust center frequency and bandwidth, the approach presented here involves offline training data, and then testing on an unseen target trajectory. By using reinforcement learning, the radar is no longer restricted to using a fixed band/group of bands; the cognitive radar learns the interference behavior and is able to predict its behavior in advance. Additionally, reinforcement learning allows the radar to achieve this behavior without having to be explicitly programmed to do so for each state, a feat that would have been very expensive in terms of hours spent in design and verification. The use of reinforcement learning frees traditional radars from the restrictions mentioned in Chapter 1, allowing it to continue target tracking functions, while coexisting in the increasingly dense radio-frequency spectrum.

The application of Markov Decision Processes to target tracking in cognitive radars was introduced in this work. The goal of this application is to apply reinforcement learning to enable the radar to maintain the target track despite the presence of interference. Specifically, MDPs are used to predict and avoid interference. The results indicate that

(1) The radar is able to learn where the interferer will be in frequency (for the interference models examined) in the next time slot, and use the contiguous bands where the interferer does not exist to maximal benefit;

(2) When the target is sufficiently close to the radar, the radar can trade SINR for increased bandwidth and still maintain positive SINR, despite using the same band(s) as the interferer;

(3) The radar reduces bandwidth usage to increase SINR as the target moves away;

(4) The radar is able to learn how often an intermittent interferer will transmit and use the bands accordingly; the radar uses all bands for the low probability case, and for the high probability case, the radar behaves similar to the constant interference case by choosing to avoid the interferer until the target is close and its SINR is sufficiently high;

(5) The radar is able to predict the band usage of a frequency-hopping communications system with short hopping patterns, and the radar is able to adjust its own bandwidth accordingly.

(6) The radar can learn where an interferer is localized, and avoid the interferer's frequency bands prior to the radar's beam entering the area with interference.

It is worthwhile to note this work demonstrates the applicability of Markov decision processes and reinforcement learning to solving this type of problem. However, there are some challenges with this approach, first of which is the state space size. If the state space is increased[1], the training process will become more complex and will take more time. To resolve this, techniques could be used to reduce the state space to a more manageable level. The other issue is this problem was modeled using *fully observable* Markov decision processes. Under full observability, what the radar observes/measures also matches the true values. When this ideal assumption is removed, we have a *partially observable* Markov decision process (POMDP), in which the radar's measurements of the environment do not necessarily coincide with the true values (which could be due to noise, for example) [9]. When the radar observes information about the environment, it doesn't know with certainty which state it is in, but rather has a set of possible states it could be in, each with an associated probability. While POMDPs may provide a more realistic model, they come at a cost of computational complexity. Therefore, techniques that can facilitate the learning process on a more complex model would be very helpful. For example, POMDPs could be transformed into a set of solvable MDPs, with one MDP per belief state. Additionally, to demonstrate the applicability of reinforcement learning, the model in this work abstracted out the actual radar signal processing. When the abstractions are removed, the model then

[1]This work investigated a band comprised of five (5) subbands; higher number of subbands increases the state space exponentially.

has to account for imperfections in the measurement of range and velocity (e.g. due to noise, or the range-Doppler coupling effect due to using an LFM waveform).

Future work will involve revisiting the assumptions discussed in Section II, and studying the effect of each. Specifically, that could entail studying

(1) An interferer that moves with respect to the radar, and the dependence of location on the received interference;

(2) Modeling an actual communications protocol, such as LTE or WiFi for the communications system;

(3) Real world experiments with cognitive radar and cognitive radio testbeds;

(4) Explicitly modeling the radar environment, atmospheric effects, multipath, clutter, and terrain;

(5) Examining the effect of an intelligent interferer;

(6) Other reinforcement learning techniques, particularly those that can reduce the state space size and training time.

Additional future work should study how to speed up the learning process by using knowledge that some transitions cannot occur. Due to target motion characteristics, the target can only transition to up to eight neighboring position states, which (in this model), rules out approximately fifty remaining position states, thereby reducing the state space.

Future work could also study incorporating received power of interference in the model. Rather than considering interference presence as a binary value on $[0, 1]$, the interval could be quantized into sub intervals, each indicating relative power of interference. For example, with four levels of interference, the interference could take any value from 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75, and 0.75 to 1. The radar could take advantage of bands that have interference, but is minimal enough to not have a severe impact on SINR. This would come at a cost of a larger state space, which would increase from $\rho\nu 2^N$ to $\rho\nu 2^{QN}$, where Q is the number of quantized levels.

Additional work could study the effect of more than one interferer in the environment. Instead of the interference occupancy vector looking like $\boldsymbol{\theta} = [0\,0\,1\,0\,0]$, it could look like $\boldsymbol{\theta} = [0\,0\,1\,1\,0]$. Part of the challenge would be simulating the performance of the radar when there are different types of interferers in the environment (*e.g.* triangle sweep and intermittent), and designing a step-frequency radar that can utilize discontinuous bands, such as when $\boldsymbol{\theta} = [1\,1\,0\,1\,0]$.

Appendix A

Analysis of Interference Cases

The notation used in the analysis is as follows: $R(\text{Action}, \text{Observed Interference})$ denotes the reward received given the action taken and the observed interference (the interference will stay constant); $R(\text{Action}, \text{Observed Interference} \rightarrow \text{Future Interference})$ is the reward received given the action taken, and observed interference, which will transition to a future interference state; $R_{L\text{B}}$, where L is the number of bands used, for example, $R_{2\text{B}}$ means the reward dealt for the radar using two bands; $R_{\text{SINR+}}$ is the worst-case reward dealt for positive SINR (which is +1); and $R_{\text{SINR-}}$ is the reward dealt for negative SINR (which is -35 for a five band scenario).

Table A.1: Value functions for high probability of transmission interference

Scenario	Value Function	
Fewer bands High SINR	$\begin{aligned} V(s) &= 0.9R([01111],[10000]) + 0.1R([01111],[00000]) \\ &= 0.9(R_{4B} + R_{SINR+}) + 0.1(R_{4B} + R_{SINR+}) \\ &= 0.9(30+1) + 0.1(30+1) \\ V(s) &= 31 \end{aligned}$	(A.1)
All bands Low SINR	$\begin{aligned} V(s) &= 0.9R([11111],[10000]) + 0.1R([11111],[00000]) \\ &= 0.9(R_{5B} + R_{SINR-}) + 0.1(R_{5B} + R_{SINR+}) \\ &= 0.9(40-35) + 0.1(40+1) = 0.9(5) + 0.1(41) \\ V(s) &= 8.6 \end{aligned}$	(A.2)
All bands High SINR	$\begin{aligned} V(s) &= 0.9R([11111],[10000]) + 0.1R([11111],[00000]) \\ &= 0.9(R_{5B} + R_{SINR+}) + 0.1(R_{5B} + R_{SINR+}) \\ &= 0.9(40+1) + 0.1(40+1) \\ V(s) &= 41 \end{aligned}$	(A.3)

Table A.2: Value functions for low probability of transmission interference

Scenario	Value Function	
Fewer bands High SINR	$\begin{aligned} V(s) &= 0.1R([01111],[10000]) + 0.9R([01111],[00000]) \\ &= 0.1(R_{4B} + R_{SINR+}) + 0.9(R_{4B} + R_{SINR+}) \\ &= 0.1(30+1) + 0.9(30+1) \\ V(s) &= 31 \end{aligned}$	(A.4)
All bands Low SINR	$\begin{aligned} V(s) &= 0.1R([11111],[10000]) + 0.9R([11111],[00000]) \\ &= 0.1(R_{5B} + R_{SINR-}) + 0.9(R_{5B} + R_{SINR+}) \\ &= 0.1(40-35) + 0.9(40+1) = 0.1(5) + 0.9(41) \\ V(s) &= 37.4 \end{aligned}$	(A.5)
All bands High SINR	$\begin{aligned} V(s) &= 0.1R([11111],[10000]) + 0.9R([11111],[00000]) \\ &= 0.1(R_{5B} + R_{SINR+}) + 0.9(R_{5B} + R_{SINR+}) \\ &= 0.1(40+1) + 0.9(40+1) \\ V(s) &= 41 \end{aligned}$	(A.6)

Table A.3: Value functions for triangular sweep interferer, without memory

Scenario	Value Function	
[00001] Using Policy	$V(s) = R([11100],[00001] \to [00010])$ $= (R_{3B} + R_{\text{SINR+}}) = 20 + 1$ $V(s) = 21$	(A.7)
[00010] Using Policy	$V(s) = 0.5R([11110],[00010] \to [00100])$ $+ 0.5R([11110],[00010] \to [00001])$ $= 0.5(R_{4B} + R_{\text{SINR-}}) + 0.5(R_{4B} + R_{\text{SINR+}})$ $= 0.5(30 - 35) + 0.5(30 + 1) = 0.5(-5) + 0.5(31)$ $V(s) = 13$	(A.8)
[00010] Using 1 Band Less Than Policy	$V(s) = 0.5R([11100],[00010] \to [00100])$ $+ 0.5R([11100],[00010] \to [00001])$ $= 0.5(R_{3B} + R_{\text{SINR-}}) + 0.5(R_{3B} + R_{\text{SINR+}})$ $= 0.5(20 - 35) + 0.5(20 + 1) = 0.5(-5) + 0.5(21)$ $V(s) = 3$	(A.9)
[00010] Completely Avoiding	$V(s) = 0.5R([11000],[00010] \to [00100])$ $+ 0.5R([11000],[00010] \to [00001])$ $= 0.5(R_{2B} + R_{\text{SINR+}}) + 0.5(R_{2B} + R_{\text{SINR+}})$ $= 0.5(10 + 1) + 0.5(10 + 1) = 0.5(11) + 0.5(11)$ $V(s) = 11$	(A.10)
[00100] Using Policy	$V(s) = 0.5R([11111],[00100] \to [01000])$ $+ 0.5R([11111],[00100] \to [00010])$ $= 0.5(R_{5B} + R_{\text{SINR-}}) + 0.5(R_{5B} + R_{\text{SINR-}})$ $= 0.5(40 - 35) + 0.5(40 - 35) = 0.5(5) + 0.5(5)$ $V(s) = 5$	(A.11)
[00100] Avoiding half the time	$V(s) = 0.5R([11100],[00100] \to [01000])$ $+ 0.5R([11100],[00100] \to [00010])$ $= 0.5(R_{3B} + R_{\text{SINR-}}) + 0.5(R_{3B} + R_{\text{SINR+}})$ $= 0.5(20 - 35) + 0.5(20 + 1) = 0.5(-15) + 0.5(21)$ $V(s) = 3$	(A.12)
[00100] Completely Avoiding	$V(s) = 0.5R([00100],[00100] \to [01000])$ $+ 0.5R([00100],[00100] \to [00010])$ $= 0.5(R_{1B} + R_{\text{SINR+}}) + 0.5(R_{1B} + R_{\text{SINR+}})$ $= 0.5(0 + 1) + 0.5(0 + 1) = 0.5(1) + 0.5(1)$ $V(s) = 1$	(A.13)

Table A.4: Value functions for triangular sweep interferer, with memory

Scenario	Value Function	
Previous:[00010] Current:[00001] Using Policy	$V(s) = R([11100],[00001] \rightarrow [00010])$ $= (R_{3\mathrm{B}} + R_{\mathrm{SINR}+}) = 20 + 1$ $V(s) = 21$	(A.14)
Previous:[00001] Current:[00010] Using Policy	$V(s) = R([11000],[00010] \rightarrow [00100])$ $= (R_{2\mathrm{B}} + R_{\mathrm{SINR}+}) = (10 + 1)$ $V(s) = 11$	(A.15)
Previous:[00010] Current:[00100] Using Policy	$V(s) = R([00111],[00100] \rightarrow [01000])$ $= (R_{3\mathrm{B}} + R_{\mathrm{SINR}+}) = (20 + 1)$ $V(s) = 21$	(A.16)
Previous: [00100] Current:[01000] Using Policy	$V(s) = R([01111],[01000] \rightarrow [10000])$ $= (R_{4\mathrm{B}} + R_{\mathrm{SINR}+}) = (30 + 1)$ $V(s) = 31$	(A.17)
Previous: [01000] Current:[10000] Using Policy	$V(s) = R([00111],[10000] \rightarrow [01000])$ $= (R_{3\mathrm{B}} + R_{\mathrm{SINR}+}) = (20 + 1)$ $V(s) = 21$	(A.18)
Previous:[10000] Current:[01000] Using Policy	$V(s) = R([00011],[01000] \rightarrow [00100])$ $= (R_{2\mathrm{B}} + R_{\mathrm{SINR}+}) = (10 + 1)$ $V(s) = 11$	(A.19)
Previous:[01000] Current:[00100] Using Policy	$V(s) = R([11100],[00100] \rightarrow [00010])$ $= (R_{3\mathrm{B}} + R_{\mathrm{SINR}+}) = (20 + 1)$ $V(s) = 21$	(A.20)
Previous:[00100] Current:[00010] Using Policy	$V(s) = R([11110],[00010] \rightarrow [00001])$ $= (R_{4\mathrm{B}} + R_{\mathrm{SINR}+}) = (30 + 1)$ $V(s) = 31$	(A.21)
Previous:[00010] Current:[00001] Using Policy	$V(s) = R([11100],[00001] \rightarrow [00010])$ $= (R_{3\mathrm{B}} + R_{\mathrm{SINR}+}) = (20 + 1)$ $V(s) = 21$	(A.22)

Appendix B

Training and Testing Algorithm

for *Each training run* **do**
    Randomly select a starting position and target velocity;
    Add "noise" to position and velocity;
    Calculate Initial SINR;
    for *Each time index of one training run* **do**
        Calculate initial state;
        Randomly select a valid action;
        Determine bandwidth used, update interference, position, range, and SINR;
        Determine new state;
        Update $\mathcal{T}$ and $\mathcal{R}$
    end
end
Using Policy Iteration, determine optimal policy;
for *Each testing run* **do**
    Using a user-defined trajectory that was not previously trained on;
    Calculate Initial SINR;
    for *Each time index* **do**
        Calculate initial state;
        Select an action from the policy;
        Determine bandwidth used, update interference, position, range, and SINR;
        Determine new state;
    end
    Create plot of Rewards, Bandwidth, SINR, Range, Actions, and Interference States
end

Algorithm 1: Algorithm for training radar and testing its performance

Bibliography

[1] H. Griffiths, L. Cohen, S. Watts, E. Mokole, C. Baker, M. Wicks, and S. Blunt, "Radar spectrum engineering and management: Technical and regulatory issues," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 85–102, Jan 2015.

[2] F. C. Commission *et al.*, "Auction of advanced wireless services (aws-3) licenses closes," *Wash. DC, DA*, pp. 15–131, 2015.

[3] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE personal communications*, vol. 6, no. 4, pp. 13–18, 1999.

[4] A. Martone, "Cognitive radar demystified," *URSI Bulletin*, no. 350, pp. 10–22, 2014.

[5] G. E. Smith, Z. Cammenga, A. Mitchell, K. L. Bell, J. Johnson, M. Rangaswamy, and C. Baker, "Experiments with cognitive radar," *IEEE Aerospace and Electronic Systems Magazine*, vol. 31, no. 12, pp. 34–46, December 2016.

[6] K. L. Bell, C. J. Baker, G. E. Smith, J. T. Johnson, and M. Rangaswamy, "Cognitive radar framework for target detection and tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, pp. 1427–1439, Dec 2015.

[7] S. Haykin, Y. Xue, and P. Setoodeh, "Cognitive radar: Step toward bridging the gap between neuroscience and engineering," *Proceedings of the IEEE*, vol. 100, no. 11, pp. 3102–3130, Nov 2012.

[8] J. M. Fuster, *Cortex and mind: Unifying cognition.* Oxford university press, 2003.

[9] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002. [Online]. Available: https://www.amazon.com/Artificial-Intelligence-Modern-Approach-2nd/dp/0137903952%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0137903952

[10] M. Levorato, S. Firouzabadi, and A. Goldsmith, "A learning framework for cognitive interference networks with partial and noisy observations," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3101–3111, September 2012.

[11] M. A. Richards, J. A. Scheer, and W. A. Holm. SciTech Publishing, 2010. [Online]. Available: http://app-knovel-com.ezproxy.lib.vt.edu/hotlink/toc/id:kpPMRVIBP8/principles-modern-radar/principles-modern-radar

[12] H. R. Raemer, *Radar systems principles*. CRC press, 1996.

[13] D. K. Barton and H. R. Ward, *Handbook of radar measurement*. Prentice Hall, 1969.

[14] S. Haykin, "Cognitive radar: a way of the future," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30–40, Jan 2006.

[15] Cognition, *Oxford English Dictionary*. Oxford University Press, 2017.

[16] W. W. Au, "A comparison of the sonar capabilities of bats and dolphins," in *Echolocation In Bats and Dolphins*, J. A. Thomas, C. Moss, and M. Vater, Eds. Chicago: The University of Chicago Press, 2004, p. xiii.

[17] J. D. Pye, *Echolocation Signals and Echoes in Air*. Boston, MA: Springer US, 1980, pp. 309–353. [Online]. Available: https://doi.org/10.1007/978-1-4684-7254-7_14

[18] W. W. Au, *The sonar of dolphins*. Springer Science & Business Media, 2012.

[19] W. W. Au and R. H. Penner, "Target detection in noise by echolocating atlantic bottlenose dolphins," *The Journal of the Acoustical Society of America*, vol. 70, no. 3, pp. 687–693, 1981.

[20] W. W. Au, P. W. Moore, and D. A. Pawloski, "Detection of complex echoes in noise by an echolocating dolphin," *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 662–668, 1988.

[21] W. W. Au and C. W. Turl, "Target detection in reverberation by an echolocating atlantic bottlenose dolphin (t ursiopstruncatus)," *The Journal of the Acoustical Society of America*, vol. 73, no. 5, pp. 1676–1681, 1983.

[22] C. Assad, B. Rasnow, and P. K. Stoddard, "Electric organ discharges and electric images during electrolocation," *Journal of Experimental Biology*, vol. 202, no. 10, pp. 1185–1193, 1999.

[23] J. Bastian and J. Yuthas, "The jamming avoidance response of eigenmannia: Properties of a diencephalic link between sensory processing and motor output," *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, vol. 154, no. 6, pp. 895–908, 1984.

[24] A. Watanabe and K. Takeda, "The change of discharge frequency by ac stimulus in a weak electric fish," *Journal of Experimental Biology*, vol. 40, no. 1, pp. 57–66, 1963.

[25] A. Martone, K. Sherbondy, K. Ranney, and T. Dogaru, "Passive sensing for adaptable radar bandwidth," in *2015 IEEE Radar Conference (RadarCon)*, May 2015, pp. 0280–0285.

[26] S. S. Bhat, R. M. Narayanan, and M. Rangaswamy, "Bandwidth sharing and scan scheduling in multimodal radar with communications and tracking," *IETE Journal of Research*, vol. 59, no. 5, pp. 551–562, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.4103/0377-2063.123761

[27] A. Martone, K. Ranney, K. Sherbondy, K. Gallagher, and S. Blunt, "Spectrum allocation for non-cooperative radar coexistence," *IEEE Transactions on Aerospace and Electronic Systems*, vol. PP, no. 99, pp. 1–1, 2017.

[28] A. Martone, K. Gallagher, K. Sherbondy, A. Hedden, and C. Dietlein, "Adaptable waveform design for enhanced detection of moving targets," *IET Radar, Sonar & Navigation*, vol. 11, no. 10, pp. 1567–1573, 2017.

[29] A. E. Mitchell, G. E. Smith, K. L. Bell, and M. Rangaswamy, "Single target tracking with distributed cognitive radar," in *2017 IEEE Radar Conference (RadarConf)*, May 2017, pp. 0285–0288.

[30] A. Basit, I. M. Qureshi, W. Khan, A. N. Malik, and B. Shoaib, "Beam pattern synthesis for a cognitive frequency diverse array radar to localize multiple targets with same direction but different ranges," in *2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Jan 2016, pp. 682–688.

[31] N. Sharaga, J. Tabrikian, and H. Messer, "Optimal cognitive beamforming for target tracking in mimo radar/sonar," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, pp. 1440–1450, Dec 2015.

[32] P. M. Djuric, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, Sep 2003.

[33] B. Z. Bobrovsky, E. Mayer-Wolf, and M. Zakai, "Some classes of global cramer-rao bounds," *Ann. Statist.*, vol. 15, no. 4, pp. 1421–1438, 12 1987. [Online]. Available: http://dx.doi.org/10.1214/aos/1176350602

[34] J. Lundén and V. Koivunen, "Deep learning for hrrp-based target recognition in multistatic radar systems," in *2016 IEEE Radar Conference (RadarConf)*, May 2016, pp. 1–6.

[35] J. Lombacher, M. Hahn, J. Dickmann, and C. Whler, "Potential of radar for static object classification using deep learning methods," in *2016 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, May 2016, pp. 1–4.

[36] A. Vasalos, N. Uzunoglu, H. G. Ryu, and I. Vasalos, "Neural network target classification for concealed weapon radar detection," in *Digital Signal Processing (DSP), 2013 18th International Conference on*, July 2013, pp. 1–6.

[37] Y. A. Nijsure, G. Kaddoum, N. K. Mallat, G. Gagnon, and F. Gagnon, "Cognitive chaotic uwb-mimo detect-avoid radar for autonomous uav navigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 11, pp. 3121–3131, Nov 2016.

[38] C. Bentes, D. Velotto, and S. Lehner, "Target classification in oceanographic sar images with deep neural networks: Architecture and initial results," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2015, pp. 3703–3706.

[39] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, "Target classification using the deep convolutional networks for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4806–4817, Aug 2016.

[40] M. D. Scherreik and B. D. Rigling, "Open set recognition for automatic target classification with rejection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 2, pp. 632–642, April 2016.

[41] F. Benedetto, F. R. Fulginei, A. Laudani, and G. Albanese, "Automatic aircraft target recognition by isar image processing based on neural classifier," 2012.

[42] *A Dictionary of Computing (Oxford Quick Reference)*. Oxford University Press, 2010. [Online]. Available: https://www.amazon.com/Dictionary-Computing-Oxford-Quick-Reference/dp/0199234000%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0199234000

[43] M. Martorella, E. Giusti, A. Capria, F. Berizzi, and B. Bates, "Automatic target recognition by means of polarimetric isar images and neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 11, pp. 3786–3794, Nov 2009.

[44] K.-T. Kim, D.-K. Seo, and H.-T. Kim, "Efficient radar target recognition using the music algorithm and invariant features," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 3, pp. 325–337, Mar 2002.

[45] Z. Zhang, S. Salous, J. Zhu, and D. Song, "A novel waveform selection method for cognitive radar during target tracking based on the wind driven optimization technique," in *IET International Radar Conference 2015*, Oct 2015, pp. 1–8.

[46] Z. Rongwen, L. Yanpeng, and J. Yafei, "Cognitive radar waveform diversity for anti-passive false target jamming in an active radar seeker," in *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, Sept 2015, pp. 1742–1745.

[47] P. Chen and L. Wu, "Waveform design for multiple extended targets in temporally correlated cognitive radar system," *IET Radar, Sonar Navigation*, vol. 10, no. 2, pp. 398–410, 2016.

[48] M. L. Manna, P. Monsurr, P. Tommasino, and A. Trifiletti, "Adaptive spectrum controlled waveforms for cognitive radar," in *2016 IEEE Radar Conference (RadarConf)*, May 2016, pp. 1–4.

[49] Y. Rufang, G. Rongbing, T. Guangfu, and H. Jie, "Range-doppler and anti-interference performance of cognitive radar detection waveform," in *2015 12th IEEE International Conference on Electronic Measurement Instruments (ICEMI)*, vol. 02, July 2015, pp. 607–612.

[50] A. Martone, D. McNamara, G. Mazzaro, and A. Hedden, *Cognitive Nonlinear Radar*, 2013.

[51] A. F. Martone, K. A. Gallagher, K. D. Sherbondy, K. I. Ranney, T. V. Dogaru, G. J. Mazzaro, and R. M. Narayanan, "Adaptable bandwidth for harmonic step-frequency radar," *International Journal of Antennas and Propagation*, vol. 2015, 2015.

[52] Z. Wang, W. Q. Wang, and J. Xiong, "Cognitive target tracking using fda radar for increased sinr performance," in *2016 IEEE Radar Conference (RadarConf)*, May 2016, pp. 1–4.

[53] W. Q. Wang, "Moving-target tracking by cognitive rf stealth radar using frequency diverse array antenna," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 3764–3773, July 2016.

[54] C. Kreucher, K. Bell, and D. Sobota, "A comparison of tracking algorithms for supermaneuverable targets," in *2015 18th International Conference on Information Fusion (Fusion)*, July 2015, pp. 534–541.

[55] K. L. Bell, J. T. Johnson, G. E. Smith, C. J. Baker, and M. Rangaswamy, "Cognitive radar for target tracking using a software defined radar system," in *2015 IEEE Radar Conference (RadarCon)*, May 2015, pp. 1394–1399.

[56] L. O. Wabeke and W. A. J. Nel, "Utilizing q-learning to allow a radar to choose its transmit frequency, adapting to its environment," in *2010 2nd International Workshop on Cognitive Information Processing*, June 2010, pp. 263–268.

[57] J. Oksanen, J. Lundn, and V. Koivunen, "Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks," *Neurocomputing*, vol. 80, pp. 102 – 110, 2012, special Issue on Machine Learning for Signal Processing 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121100600X

[58] A. Kolobov, "Planning with markov decision processes: An ai perspective," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–210, 2012.

[59] M. van Otterlo and M. Wiering, *Reinforcement Learning and Markov Decision Processes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 3–42. [Online]. Available: https://doi.org/10.1007/978-3-642-27645-3_1

[60] O. Ibe, *Fundamentals of applied probability and random processes*. Academic Press, 2014.

[61] S. Haykin and J. M. Fuster, "On cognitive dynamic systems: Cognitive neuroscience and engineering learning from each other," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 608–628, 2014.

[62] I. Chadès, G. Chapron, M.-J. Cros, F. Garcia, and R. Sabbadin, "Markov decision processes (mdp) toolbox," Jan 2015, http://www7.inra.fr/mia/T/MDPtoolbox/MDPtoolbox.html. [Online]. Available: http://www7.inra.fr/mia/T/MDPtoolbox/MDPtoolbox.html

[63] J. Wintenby and V. Krishnamurthy, "Hierarchical resource management in adaptive airborne surveillance radars," *IEEE Transactions on Aerospace and Electronic systems*, vol. 42, no. 2, pp. 401–420, 2006.

[64] J. Wintenby, *Resource allocation in airborne surveillance radar*. Chalmers University of Technology, 2003.

[65] Y. Li, L. W. Krakow, E. K. Chong, and K. N. Groom, "Approximate stochastic dynamic programming for sensor scheduling to track multiple targets," *Digital Signal Processing*, vol. 19, no. 6, pp. 978–989, 2009.

[66] B. La Scala, W. Moran, and R. Evans, "Optimal adaptive waveform selection for target detection," in *Radar Conference, 2003. Proceedings of the International*. IEEE, 2003, pp. 492–496.

[67] B. F. La Scala and B. Moran, "Optimal target tracking with restless bandits," *Digital Signal Processing*, vol. 16, no. 5, pp. 479–487, 2006.