3-2022

Team Air Combat using Model-based Reinforcement Learning

David A. Mottice

TEAM AIR COMBAT USING MODEL-BASED REINFORCEMENT LEARNING

THESIS

Mottice, David, A, 2d Lt, USAF

AFIT-ENS-MS-22-M-157

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-22-M-157

TEAM AIR COMBAT USING MODEL-BASED REINFORCEMENT LEARNING

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Mottice, David, A, BS

2d Lt, USAF

March 24, 2022

AFIT-ENS-MS-22-M-157

TEAM AIR COMBAT USING MODEL-BASED REINFORCEMENT LEARNING

THESIS

Mottice, David, A, BS
2d Lt, USAF

Committee Membership:

Matthew Robbins, PhD
Advisor

Maj Phillip Jenkins, PhD
Reader

AFIT-ENS-MS-22-M-157

Abstract

Executing within visual range air combat requires a pilot to make numerous interconnected decisions each second while flying at speeds close to Mach 1. Fighter pilots spend years in pilot training learning the tactics to be successful in these engagements. However, the speed and quality of their decision making is limited by human biology. The emergence of autonomous unmanned combat aerial vehicles (AUCAVs) exploits this limitation and changes the fundamentals of air combat. However, recent research focuses on one-versus-one engagements and ignores a fundamental rule of air combat – never fly alone. We formulate the first generalized air combat maneuvering problem (ACMP), called the MvN ACMP, wherein M friendly AUCAVs engage against N enemy AUCAVs, developing a Markov decision process (MDP) model to control the team of M Blue AUCAVs. The MDP model leverages a 5-degree-of-freedom aircraft state transition model and formulates a directed energy weapon capability. The continuous and high dimensional nature of the state space prevents the use of classical dynamic programming solution approaches to determine optimal policies. Instead, an approximate dynamic programming (ADP) approach is adopted wherein an approximate policy iteration algorithmic strategy is implemented to attain high-quality approximate policies relative to a high performing benchmark policy. The ADP algorithm utilizes a multi-layer neural network for the value function approximation regression mechanism. One-versus-one and two-versus-one scenarios are constructed to test whether an AUCAV can outmaneuver and destroy a superior enemy AUCAV. The performance is evaluated across offensive, defensive, and neutral starts, leading to six problem instances. The ADP policies outperform the position-energy benchmark policy in four of six problem instances. Results show the ADP approach mimics

certain basic fighter maneuvers and section tactics.

Table of Contents

Page

Abstract iv

List of Figures viii

List of Tables x

I. Introduction 1

II. Literature Review 6

2.1 Air Combat 6
2.1.1 Directed Energy Weapons 9
2.2 Air Combat Maneuvering Problem 14
2.2.1 Early Attempts 14
2.2.2 Model-Based RL 15
2.3 Single Agent to Multi-Agent 19
2.3.1 Multi-Agent RL 19
2.3.2 Multi-Agent Air Combat 23
2.4 Solution Approaches 27
2.4.1 Approximate Dynamic Programming 27
2.4.2 Reinforcement Learning 28

III. Methodology 30

3.1 Problem Description 30
3.2 MDP Formulation 32
3.2.1 Agent space 32
3.2.2 State space 32
3.2.3 Action space 34
3.2.4 Transitions 37
3.2.5 Contributions 43
3.2.6 Objective function and Bellman equation 44
3.3 Model-based Reinforcement Learning Approach 45
3.3.1 Basis Functions 46
3.3.2 Exploration Mechanism 47
3.3.3 Sparse Reward Mechanism 47
3.3.4 Algorithm Formulation 49

IV. Testing, Results, and Analysis 55

4.1 Representative Scenario Description 55
4.2 Benchmark Policy 56
4.3 1v1 Designed Experiment and Analysis 58

Page

4.3.1 Offensive Starts 61
4.3.2 Defensive Starts 66
4.3.3 Neutral Starts 71
4.4 2v1 Designed Experiment and Analysis 75
4.4.1 Offensive Starts 79
4.4.2 Defensive Starts 80
4.4.3 Neutral Starts 83

V. Conclusions and Recommendations 87

Bibliography 90

List of Figures

Figure		Page
1	Standard Missile Components (Shaw, 1985)	8
2	Gaussian Intensity Profile (Zohuri, 2016)	12
3	Propagation of a Laser Beam (Zohuri, 2016)	12
4	Low Atmosphere Model Example (Zohuri, 2016)	13
5	Seven Basic Flight Maneuvers (Fang et al., 2016)	17
6	Axes of Team Coordination (Toubman, 2020)	25
7	Effect of β_p on N_k^H where $L = 20,000$	49
8	Adaptive State Sampling Scheme, $k = 1$	50
9	Adaptive State Sampling Scheme, $k = 3$	50
10	Adaptive State Sampling Scheme, $k = 10$	50
11	Offensive Starts	56
12	Defensive Starts	56
13	Neutral Starts	56
14	1v1 Offensive Engagement in xyz Plane	64
15	1v1 Offensive Engagement in xy Plane	64
16	1v1 Offensive Engagement in xz Plane	64
17	1v1 Offensive Engagement: p_{kill} Trends	65
18	1v1 Offensive Engagement: Angle Trade-offs	65
19	1v1 Defensive Engagement in xyz Plane	67
20	1v1 Defensive Engagement in xy Plane	67
21	1v1 Defensive Engagement in xz Plane	67
22	1v1 Defensive Engagement: Angle Trade-off	69

Figure Page

23 1v1 Defensive Engagement: p_{kill} Trends 70

24 1v1 Neutral Engagement in xyz Plane 73

25 1v1 Neutral Engagement in xy Plane 73

26 1v1 Neutral Engagement in yz Plane 73

27 Position-Only Scores for Red AUCAV 1v1 Neutral Engagement 74

28 1v1 Neutral Engagement: Angle Trade-off 75

29 2v1 Defensive Engagement in xyz Plane 81

30 2v1 Defensive Engagement: Angle Trade-offs 81

31 2v1 Neutral Engagement in xyz Plane 84

32 2v1 Neutral Engagement in xz Plane 84

33 1v1 Neutral Engagement: Angle Trade-offs 84

34 The Half-Split (Shaw, 1985) 85

List of Tables

Table		Page
1	1v1 Factors and Levels	59
2	1v1 Designed Experiment Top Results	60
3	1v1 Overall Results	60
4	1v1 Offensive Performance over 100 Replications	61
5	1v1 Defensive Performance over 100 Replications	66
6	1v1 Neutral Performance over 100 Replications	71
7	2v1 Factors and Levels	76
8	2v1 Designed Experiment Top Results	78
9	2v1 Overall Results	79
10	2v1 Offensive Performance over 20 Replications	79
11	2v1 Defensive Performance over 20 Replications	80
12	2v1 Neutral Performance over 20 Replications	83

I. Introduction

The ultimate advantage for any military is airpower, which is the ability to control and attack through the air with little resistance (United States Department of the Air Force, 2021). The role of the Air Force is to deliver this airpower anytime and anywhere. However, an increase in global disorder, combined with the rapid development of technology, is redefining how and where the Air Force applies airpower. In 1947, the Air Force was created to organize, train, and equip forces primarily for prompt and sustained offensive and defensive air operations (United States Department of the Air Force, 2021). Today, the Air Force is responsible for operations across the air, space, and cyberspace domains (United States Department of the Air Force, 2021). The landscape for war is ever changing, and the reemergence of inter-state competition and the democratization of new technologies are changing this landscape quicker than ever before. Mazarr et al. (2018) notes this reemergence is the result of states either pursuing narrow goals, such as the survival of current regime, or competing for broader objectives, such as global status or a global agenda. A case of particular interest is a state pushing its own global agenda due to disagreements with current world power's global agenda, like China and the United States (US).

The growing inter-state competition is reinforced in the National Defense Strategy (NDS), when it asserts the competitive military advantage of the US has been decreasing in a world marked by increasing global disorder (United States Department of Defense, 2018). A specific instance of the growing inter-state competition is the emerging power of China and how it is creating a two-country global power

structure led by China and the US that has not been seen before in history (Mazarr et al., 2018). Senior defense leaders initiated a fundamental shift by declaring the inter-state strategic competition as the primary concern for US national security. No longer is terrorism the primary threat to US national security. For the past thirty years, the Air Force has trained air, space, and cyberspace forces for irregular warfare to combat terrorism. Today, it is training air, space, and cyberspace forces for a possible peer to peer conflict. However, the rapid development of emerging technologies erodes any notion of certainty on what this conflict will look like.

Senior defense leaders recognize this erosion and are investing in various areas of research to modernize key capabilities across the Department of Defense (DOD), with *advanced autonomous systems* as one of the key areas. Various leaders, subject matter experts, and authors compare the modern race in military artificial intelligence (AI) to the development of atomic bombs and stealth aircraft. It is true that AI will transform the way the US military fights war. However, Morgan et al. (2020) describe a key distinction between such a comparison, noting the US military does not enjoy a monopoly or even a first mover advantage in the competition of AI. Unlike the development of atomic weapons and stealth aircraft, the US military has and will face significant international competition in AI from the onset. Recently, small autonomous drones identified a retreating enemy force and dive-bombed trucks and individual soldiers, causing mass panic and confusion (Choudhury et al., 2019). These small autonomous drones are built by open source image recognition and autonomous flight software, demonstrating the low barrier to entry for small militias and governments to apply AI to the battlefield. Military AI is no longer science fiction; it is a reality and can be developed by almost anyone.

In the US, these autonomous systems are expected to operate in Joint All-Domain Operations (JADO) where actions by the joint force are integrated in planning and

synchronized in execution at speed and scale. In 2020, General David Goldfein, former Chief of Staff of the Air Force (CSAF), summarized JADO as the action of "turning large amounts of multi-source data into actionable intelligence" at a speed and quality that is quicker and more accurate than our enemy. In response, the Air Force Office of Scientific Research now calls for research into "rapid, effective decision making" that seeks to accelerate decisions "from the cockpit to the headquarters." In the cockpit, this speed and quality of decision making is currently constrained by the human fighter pilot.

Consider how fighter pilots can handle 10 gravitational (G) forces in a sustained turn, and recent research shows the conscious mind is capable of processing approximately 50 bits per second despite sensing 11 million bits per second (Wiliam, 2006). Then, consider these insights in the context of a within visual range (WVR) air combat scenario, commonly called the dogfight. Fighter pilots must make decisions at the speed of sound while simultaneously sifting through billions of inputs per second. Researchers believe there is an untapped potential in WVR air combat capabilities that can be uncovered if these human imposed constraints are relaxed. This relaxation will result in the autonomous unmanned combat aerial vehicle (AUCAV). Combine this improvement with the reduced training and maintenance costs relative to manned aircraft, and the AUCAV will change the fundamentals of air warfare. Therefore, the first country to develop and field an effective AUCAV to the battlefield will likely hold air superiority in the next conflict.

Significant research has been done on how to imbue a single AUCAV with the necessary intelligence to execute air combat tactics successfully. Rules-based logic is a popular choice for simulators that are currently used in fighter pilot training, where the human evolution of WVR air combat is programmed into a series of logical statements. This approach often fails to elicit intelligent behavior from an AUCAV

in a simulated scenario. Also, in the context of a real air engagement, a rules-based logic would be a security risk because the pre-determined nature of the algorithm renders the technology useless once an enemy combatant reverse engineers and learns the logic. A better, more adaptive solution approach is needed.

Model-based reinforcement learning (RL) could be the more adaptive approach. Sometimes referred to as approximate dynamic programming (ADP), model-based RL leverages a Markov decision process (MDP) model to help the AUCAV learn in a sample efficient manner. Policies provide high-quality decisions given the state of the air engagement. Calculated offline, these policies can yield near-real time decision making for an AUCAV. This approach generates intelligent AUCAV behavior in the one versus one (1v1) air combat maneuvering problem (ACMP). McGrew et al. (2010) first demonstrated the feasibility of applying model-based RL to the ACMP. This research extends the ACMP model to a M versus N (MvN) model wherein M Blue AUCAVs execute a series of actions against N Red AUCAVs, with an initial focus on the 2v1 case.

In particular, this research tests the MvN model by exploring one versus one (1v1) and two versus one (2v1) engagements. The 1v1 ACMP examines where a single friendly (i.e., Blue) AUCAV must execute a series of actions against a single adversarial (i.e., Red) AUCAV. The 2v1 ACMP wherein a team of two friendly (i.e., Blue) AUCAVs must coordinate and execute a series of actions against a single adversarial (i.e., Red) AUCAV, called section tactics, and shoot it down (Shaw, 1985). Conversely, the Red AUCAV attempts to position itself behind a Blue AUCAV and shoot it down. Both the 1v1 and 2v1 ACMP are complex and dynamic problems. It is practically impossible to mathematically model the minutiae of each interaction. Therefore, a couple key assumptions are made. First, perfect information between all AUCAVs in the engagement is assumed, where each AUCAV knows the exact

location of the other AUCAVs. This is realistic given the extensive development of sensors in recent years. Second, perfect control is assumed for each AUCAV, where each desired kinematic action is implemented with certainty.

An MDP model is formulated to characterize the MvN ACMP. This model leverages the 5 degree of freedom (5DOF) kinematic model, as formulated by Crumpacker (2021), extending it to the multi-AUCAV context. The continuous state space renders traditional, exact solution methodologies computationally intractable. Instead, an ADP solution procedure is developed. However, the addition of an AUCAV introduces added complexity. With the additional AUCAV, how do we inculcate this team of AUCAVs with the necessary AI to maneuver and shoot down an enemy AUCAV? Will behaviors change when there is full communication versus no communication between AUCAVs? What if we relax the human imposed constraints? What new tactics do we see?

The developed MDP formulation helps us analyze a simple 2v1 engagement to understand the implications and unknowns generated by extending to the multi-AUCAV context. An ADP solution procedure is developed and compared to a benchmark policy found in technical and general air combat literature to determine the efficacy of the ADP solution relative to current practices, both in simulators and ones currently taught in pilot training.

The remainder of this thesis is structured in the following manner. Chapter II overviews the existing research related to aerodynamics, 1v1 ACMP, 2v1 ACMP, and multi-agent RL (MARL). Chapter III characterizes the MDP formulation for the MvN ACMP and the ADP/MARL solution procedure leveraged to find high-quality maneuver policies for the team of Blue AUCAVs. Chapter IV explores a set of quantitative tests where the model and solution approach are explored. Chapter V finishes and discusses future vectors of this research.

II. Literature Review

This research focuses primarily on developing an appropriate model and solution approach to adequately characterize the MvN WVR ACMP and produce high-quality policies, respectively. The goal is to elicit intelligent team behavior. Secondarily, this research explores whether and to what degree of communication is needed within the team and how the addition of emerging technologies, such as directed energy weapons, affect the WVR ACMP. Four areas of scientific literature are explored. First, the publicly available collection of theories and tactics on air combat is studied to provide context and domain knowledge. Second, previous approaches to variants of the ACMP are explored to assist in characterizing the 2v1 ACMP and generate different problem features. Third, the leap from controlling a single agent to multiple agents is explored to inform the additional modeling and solution approach challenges. Fourth, various MARL algorithms are analyzed to inform which solution approach is best suited for the 2v1 WVR ACMP.

2.1 Air Combat

Shaw (1985) simplifies air combat to two high-level systems: the aircraft and the mounted weapon system. The aircraft is considered the weapons platform that brings the mounted weapon system into position for firing. Meanwhile, the mounted weapon system is the weapons platform that accurately brings damage from the position of firing to the enemy combatant. The unique combination of these two high-level systems requires unique air combat strategies to be successful according to Shaw (1985). For example, an F-22 Raptor employing the standard air-to-air gun will behave differently than an F-22 Raptor employing a next generation directed energy weapon. Another notional example is an F-22 Raptor employing the standard air-to-

air gun will behave differently than an F-16 also employing a gun. Each combination of aircraft and mounted weapon system requires a unique strategy, or policy, in the context of the ACMP.

These high-level problem features do not lose their relevance when characterizing the 2v1 WVR ACMP. The aircraft, which refers to the AUCAV's kinematic model and given flight dynamics, has been extensively researched in ACMP literature, as explored in the next section. However, most of the existing research employs the standard air-to-air gun when considering the mounted weapon system. This is not an unrealistic weapon system to be employed because of the savings in cost and space on the aircraft relative to other weapons. Combined with the ability to easily define a forward weapon engagement zone (WEZ) mathematically, researchers are able to focus on improving the kinematic models in the ACMP quickly and subsequently determine high-quality policies that maneuver a Blue AUCAV directly behind the Red AUCAV before shooting with a gun.

However, many fighter pilots note that American planes have not engaged in that form of air combat in over 40 years, with some calling it an arcane way of thinking. While many pilots are emphasizing the increasing role of beyond visual range (BVR) engagements, they are also emphasizing the use of emerging weapon systems such as the F-35's ability to fire off-boresight, which effectively widens the WEZ to 180 degree range from the front of the AUCAV. Advancements like these are what is eroding the arcane way of thinking about air combat. Therefore, future research concerning the ACMP should explore the effect of the various weapon systems.

The next candidate for a weapon system is guided missiles, which are categorized by mission type. In air-to-air combat, the guided missile is an air-to-air missile (AAM). Specifically, in a WVR engagement, the short range AAM (SRAAM) is most often used. Shaw (1985) details the six common components to a successful guided

missile as 1) missile propulsion, 2) missile control, 3) missile guidance, 4) missile seekers, 5) missile fuse, and 6) missile warhead. Figure 1 shows these components in a standard missile. Missile propulsion refers to the mechanism that accelerates the missile from rest to the velocity needed to overtake and detonate near the desired target, an enemy aircraft. This typically reduces to rocket versus jet powered. In SRAAMs, rockets are the default. Missile control refers to the system responsible for maneuvering the missile in response to inputs from the guidance system, which is similar to the basic aircraft controls. These maneuvers are traditionally executed via thrust-vector control where the direction of the exhaust gases is altered to change by swiveling nozzles.

Missile guidance refers to the system that provides inputs to the missile control system. Shaw (1985) identifies four traditional guidance systems for AAMs: preset, command, beam-rider, and homing. The most effective and common is homing, where the emissions of the target are the main source of information. Missile seeker refers to the system responsible for sensing and tracking the target, which then provides the necessary information to the missile guidance. Modern SRAAMs typically employ some variant of a heat seeker, where a device contains a material that is sensitive to heat (i.e., infrared (IR) radiation). The primary external source of heat is typically

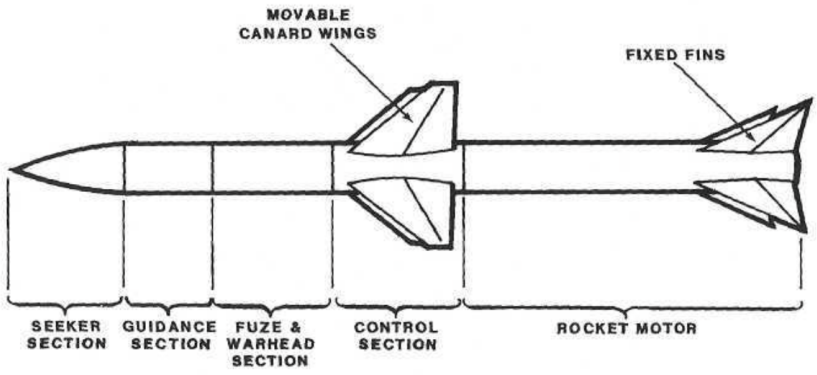

Figure 1. Standard Missile Components (Shaw, 1985)

the target's propulsion system. In context, missile seekers detect the target, missile guidance is the plan to get to the target, and missile control is the execution of that plan.

The missile WEZ is different than the gun WEZ in two ways. First, the gun WEZ is very small relative to the missile WEZ. This is due to the unavoidable limiting factors of the gun WEZ being gravity, which pulls bullets down to earth at a certain range. This leads to the need for an aircraft to be pointing directly at the target, leading to a narrow, short WEZ. Conversely, a missile WEZ leverages a much higher maximum range and does not require being directly pointed at the target, thanks to rocket power and homing systems, respectively. Second, the gun WEZ possesses an exclusive area where the missile WEZ cannot operate, typically within a thousand feet. This is due to the limiting factor of missiles being their minimum range in a WVR engagement. If the target is not maneuvering, the missile can be employed at a minimum range similar to guns. However, the focus of this research is on targets that are maneuvering. In this case, the minimum range increases because the missile must perform a high G turn to hit the target. Although modern missiles can pull over 30 Gs, if the target turns tight enough, missiles will still fail to intercept. This leads to many missiles not being employed within certain ranges.

2.1.1 Directed Energy Weapons

Directed energy weapons (DEWs) are often perceived as the weapon system of the future. The most common form is called light amplification by stimulated emission of radiation (i.e., laser), which amplifies a narrow, intense beam of coherent light. However, laser, and all other DEW, depend on energy. An important consideration and key understanding for the deployment of DEWs is how they propagate their energy. In this propagation, some energy is lost. Therefore, a DEW must propagate

more energy than needed to damage the target.

Zohuri (2016) represents the entire sequence of employing a DEW as

$$E = A\big(I \times t\big), \tag{1}$$

where E is the absorbed energy by the target, A is some fraction value to represent the amount of energy not lost in the propagation process, I is the intensity of the laser beam, and t is the emission duration of the laser on the target. Although equation (1) over-simplifies the process, it does provide a framework where assumptions can be gradually relaxed until an appropriate DEW is modeled.

When constructing a DEW, it is common to first think about how much energy is needed to destroy the target, E, which typically implies determining what kind of damage is desired. For example, the two extremes are "soft kill" or "hard kill" where the internal computer and mechanics of the target is disrupted or the target is vaporized, respectively. Hard kills will likely be the target of future research in the ACMP due to the physics and theory of a soft kill being much more nuanced. Zohuri (2016) notes that 700 J/cm^2 will punch through the frame of an airplane while the realistic energy level need to disable the entire aircraft may be five to ten times higher. This implies a successful DEW must deliver 5000 to 10000 J/cm^2 onto the target.

A fundamental concept for DEWs is intensity, which is a density function to determine how much power is being applied to some unit of area. A common metric to compare different various DEWs is peak intensity, I_p, calculated by

$$I_p = \frac{P_0 \pi}{A}, \tag{2}$$

where P_0 is the output power of the beam and A is the "beamed" area of the target.

As Equation (2) shows, the objective is to maximize the output power of the beam while minimizing the area on the target our beam hits. While it is easier to think of intensity as a scalar, it actually represents a distribution, as visualized in Figure 2, where the intensity follows a Gaussian distribution relative to the distance from the center of the beam.

The multi-objective problem of maximizing power while minimizing area on target is complicated once the atmosphere is considered (Zohuri, 2016). In a vacuum, a laser beam will naturally diverge in a linear manner, as shown in Figure 3. This phenomenon is typically called *beam spread*. However, when a laser is propagated in the atmosphere, beam spread will diverge at a faster rate according to *extinction*, which is the reduction or attenuation in the amount of radiation passing through the atmosphere. Note that extinction is the term used to group the atmospheric effects of absorption and scattering into one topic. *Absorption* is the manner by which the gaseous molecules of the atmosphere convert some photons of radiation in the laser beam to kinetic energy for the molecules. In short, absorption is the loss of energy to the atmosphere, where the atmosphere around the beam is heated (due to the increase in kinetic energy). *Scattering* occurs in the visible and IR wavelength ranges when the radiation propagates through certain air molecules and particles. Various research has been conducted into determining these coefficients at various altitudes and weather conditions. Programs such as LOWTRAN, FASCODE, MODTRAN, HITRAN, and PcLnWin can all be used to estimate these coefficients. However, access to these programs is typically restricted to appropriate government organizations and large private companies. Also, we consider the additional computational burden of calling these programs not worth the more precise estimates in large part to this topic not being the primary focus of this thesis.

Therefore, to create a model of the intensity of a beam, we need an equation

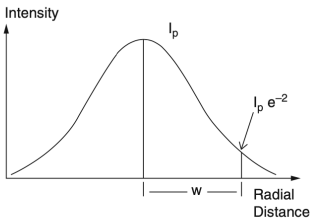

Figure 2. Gaussian Intensity Profile (Zohuri, 2016)

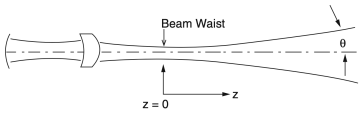

Figure 3. Propagation of a Laser Beam (Zohuri, 2016)

characterizing the intensity of a beam at some distance. Zohuri (2016) calculates the intensity of the beam waist at some distance z from the beam waist $I(z)$ as

$$I(z) = I(0)e^{-Kz}, \tag{3}$$

where $I(0)$ is the intensity at the beam waist and K is the extinction coefficient. Note the *beam waist* is the point at which the radius of the beam is smallest, and is typically denoted by $z = 0$, as shown in Figure 3. Equation 3 exhibits a few shortcomings, the major one being that we are assuming a constant K. The reality is that K changes as a function of altitude and would thus change at each point along the trajectory from $z = 0$ to the target.

Zohuri (2016) improves upon Equation (3) to yield

$$I(z') = I(0)e^{-\int_{z=0}^{z=z'} K(z)dz}, \tag{4}$$

where we *integrate* K over the path from $z = 0$ to $z = z$. This essentially sums the optical depths over each small path segment. The programs such as LOWTRAN and MODTRAN can do this with high precision. However, we can create simple model that is reasonably accurate and can be solved analytically (Zohuri, 2016).

The following model is valid for the low atmosphere, which is up to 120 kilometers, or around $393,700$ feet, from sea level. Therefore, this is can be used for the entire

area of responsibility (AOR) in our model. Zohuri (2016) uses an example of firing a laser into the air from the ground to establish this model. Figure 4 represents this scenario.

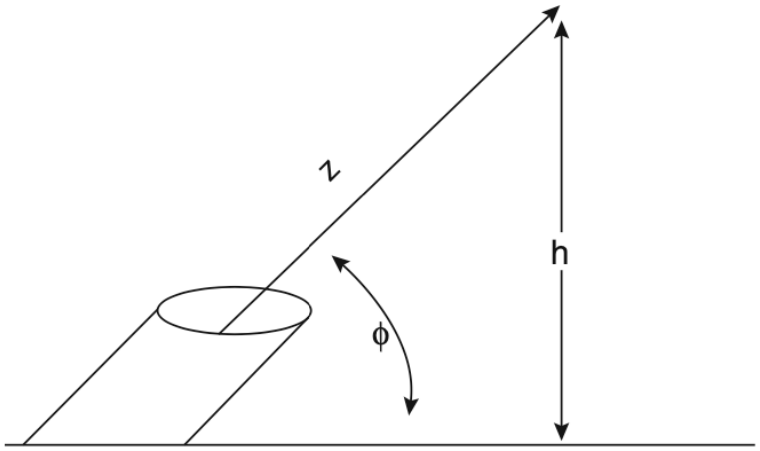

Figure 4. Low Atmosphere Model Example (Zohuri, 2016)

The extinction coefficient as a function of altitude can be approximated by

$$K(h) = K(0)e^{-\frac{h}{h_0}}, \tag{5}$$

where $h_0 \approx$ 7-km, which is an average for various molecules present at sea level. Then, supposing we fire a laser into the air at some angle ϕ, we can then relate the range z to altitude using $h = z \sin \phi$.

We can modify the above use case for the intended purpose of a laser that is flying below some AUCAV at altitude h that is being fired to some target at altitude h_t. First, we will consider the case where $h_t > h$, i.e. the target is above the AUCAV. This is similar to Figure 4 except the vertical bar is now $h_t - h$ since we are firing from h to h_t instead of 0 to h. This means $K(h_t) = K(h)e^{-\frac{h_t}{h}}$.

2.2 Air Combat Maneuvering Problem

The air combat maneuvering problem (ACMP) seeks to determine the sequence of maneuvers needed to place oneself in a position of advantage or escape from a position of disadvantage in offensive and defensive scenarios respectively, given the mounted weapon system. Top Gun, the famous 1986 movie about Naval Fighter Weapons School, popularized this problem in a cinematic setting. Various scientific communities explored the feasibility of automating this problem.

2.2.1 Early Attempts

Burgin and Owens (1975) publish the first documented research attempting to automate air combat. Decision Science Inc., a NASA contractor at the time, formulates, describes, and implements a novel technique for air combat simulation, the Adaptive Maneuvering Logic (AML). AML encodes the entire evolution of human intelligence, up to 1975, on air combat into a series of logical statements, creating a rules-based heuristic. A real time version was then implemented against human pilots in the Langley Differential Maneuvering Simulator (DMS). The results were then graphed by hand to demonstrate the successful generation of air combat behavior while being unsuccessful in shooting down human pilots.

Burgin and Sidor (1988) update AML by considering the additions of human intuition gained in the 13-year period since the first version of AML. However, Burgin and Owens (1975) and Burgin and Sidor (1988) both note that performance in air combat via a rules-based heuristic is capped to the performance level of human pilots and the existing knowledge bank of air combat tactics. Also, if an enemy combatant can train to learn the rules-based heuristic and existing knowledge bank, then the AUCAV would be ineffective in combat. Therefore, a more adaptive solution is needed.

Virtanen et al. (2006) attempt to produce a more adaptive solution by modeling the ACMP as an influence diagram game. An influence diagram is a directed acyclic graph (DAG) to describe a decision making process and provides a methodology to rank decision alternatives. The authors combine influence diagrams with the theory of dynamic programming, made popular by Bellman (1952). This approach is successful; however, it is restricted to a limited planning horizon due to computational complexity, even as technology increases computational power.

2.2.2 Model-Based RL

The 2v1 WVR ACMP can be viewed as a sequential decision-making process wherein a series of maneuvers must be executed to attain and maintain a position of advantage. The natural field for this type of problem is operations research, where analytical approaches are applied to improve decision making. In operations research, a Markov decision process (MDP) model is the common mathematical framework applied to analyze a sequential decision-making problem under uncertainty.

MDP models are characterized by five components (Puterman, 1994). First, a set of decision epochs is defined, typically denoted as $\mathcal{T}$. Each element within this set represents a point in time at which decisions are made. This set can be finite or infinite and continuous or discrete based on the system of interest. Second, a set of system states is defined, typically denoted as S. This set must contain all possible states the system of interest can occupy. At each decision epoch, the system occupies some state $s \in S$. Third, a set of available actions is defined, typically denoted as A_s. At each decision epoch, the decision maker (or agent) chooses an action from the set of allowable actions given the current state. Fourth, a reward (or cost) structure is defined based on the state-action pair of the system, typically denoted as $C(\cdot)$. Careful consideration must be given to this structure as this is the

main factor in what actions are considered optimal. Fifth, and finally, a transition dynamics function, $T(s_{t+1}|s_t, a_t)$ is defined to map each state-action pair at epoch t to a distribution of states at epoch $t + 1$.

MDP models provide an adaptive solution to be successful in the ACMP. However, McGrew et al. (2010) demonstrate solving for optimal policies is computationally intractable, even in the simplest representations of the ACMP. The authors consider a two-dimensional engagement where altitude is not considered, and velocity is fixed. The set of system states is defined as the absolute position and orientation of both Blue and Red AUCAVs. The set of available actions is simply whether the Blue AUCAV rolls right, maintains current bank angle, or rolls left. Despite the simplicity, the authors had to apply approximate dynamic programming (ADP) to yield high-quality policies. This work demonstrates the feasibility and possible success of the hybrid MDP and ADP approach in the ACMP while also highlighting the need for future research to focus on algorithmic design to be able to solve more realistic representations of the ACMP.

Fang et al. (2016) improve upon McGrew et al. (2010) by considering altitude and increasing the action space. The set of system states is still defined as the absolute position and orientation of both Blue and Red AUCAVs; however, the action space now includes the seven basic flight maneuvers (BFMs) defined by NASA: 1) maintain, 2) pull-up, 3) dive, 4) left bank, 5) right bank, 6) acceleration, and 7) deceleration. Figure 5 visualizes these maneuvers. One of the more salient contributions is the use of aircraft energy in the reward function. Shaw (1985) reinforces the importance of energy in air combat by noting that the aircraft with higher energy has an advantage as pilots can learn to trade that energy for position.

Wang et al. (2020) improve in the kinematic modeling of AUCAVs in the ACMP by redefining the set of possible actions. McGrew et al. (2010) considered three

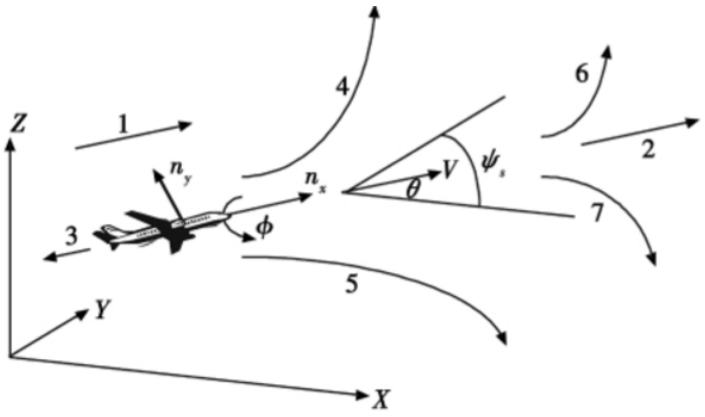

Figure 5. Seven Basic Flight Maneuvers (Fang et al., 2016)

possible actions: roll left, maintain bank angle, or roll right. Fang et al. (2016) considered seven BFMs defined by NASA and visualized in Figure 5. However, both of these approaches utilize actions that are outcomes of smaller, more well-defined actions. For example, when rolling left, or executing a left bank, the pilot must adjust the velocity, bank angle, and heading angle incrementally. Wang et al. (2020) model these smaller, well-defined actions by the set of actions being various discrete levels of velocity, bank angle, and heading angle. This tuple of actions is discrete to maintain computational tractability. The result is smoother, more realistic flight behavior from the Blue AUCAV in the engagement with the Red AUCAV.

Crumpacker (2021) contributes the most recent work with the formulation of a 5DOF kinematic model. The author extends Fang et al. (2016) by incorporating angle of attack, roll angle, throttle setting, mass and fuel remaining of both the Blue and Red AUCAV into the state space. The higher-fidelity model also incorporates information about the Blue AUCAV's gun, implementing the decision to shoot as part of the action space. The expansion of the state space yields high-quality policies that compete with and sometimes outperform the position only and position-energy benchmark. Crumpacker (2021) also notes the already high performance of the position and position-energy benchmarks, which make great candidates to be applied to

either the Red AUCAV, or the other Blue AUCAV in the event only 1 is being controlled in the 2v1 WVR ACMP. A salient result is the emergence of realistic behavior by the Blue AUCAV, such as the wing-over and vertical loop maneuver in the 1v1 engagements.

Much of the recent ACMP research also originates from the computer science community where the ACMP is viewed from the lens of reinforcement learning (RL). RL is the goal-directed learning of an agent through interaction with its environment. Interestingly, ADP and model-based RL address the same underlying issue: planning the highest-quality set of decisions to be successful in a given scenario (Sutton and Barto, 2018). The two use different notations and terminology but perform the same function of approximating a value function, in most cases, to produce high-quality policies in an efficient manner.

Pope et al. (2021) note the complexity of the ACMP and introduce hierarchical RL as a solution approach. The traditional human approach to a complex task is to divide it into a smaller subset of more reasonable tasks. Hierarchical RL mimics this by treating the ACMP as a semi-Markov decision process (SMDP). The goal is to define macro-actions, called routines, to model the agent at different levels of temporal abstraction. The advantages of hierarchical RL over traditional methods include transfer learning, scalability, and generalization. Transfer learning and scalability assist in combating the curse of dimensionality by leveraging learned skills in new tasks and decomposing the problem into sub-problems, respectively. The authors model an agent with a 2-layer hierarchy of policies. The lower layer of policies is trained to excel in a particular region of the state space while the higher layer contains a single policy selecting which low-level policy to activate given the state of the engagement. The soft actor-critic (SAC) algorithm is leveraged to train the three low-level policies: control zone (CZ), aggressive shooter (AS), and conservative shooter

(CS). The CZ policy incentivizes maneuvers to attain and maintain a position of advantage over the Red AUCAV. The AS policy incentivizes risky (i.e., high reward, low probability) shots, from the side and head on – think high risk and high reward. The CS policy incentivizes the opposite. At the high level, a different SAC algorithm selects which low-level policy to implement given the state of the engagement. The resulting agent outperforms all defined benchmarks, finishing second at the Alpha Dogfight Trials. The agent also defeated a USAF Weapons Instructor Pilot five times in a row. Hierarchical RL exhibits promise in the 2v1 WVR ACMP, particularly in its ability to combat the curse of dimensionality. Up to this point, all research discussed in this paper is focused on the design and implementation of a single AUCAV against an adversary. The next section examines research considering the control of multiple AUCAVs.

2.3 Single Agent to Multi-Agent

The recent advancements in deep RL are in the limelight for accomplishing tasks such as defeating the best Go player in the world (Wang et al., 2016), self-driving cars (Fayjie et al., 2018), and smart home automation (Yu et al., 2019), among many others. However, many basic tasks common to human intelligence elude a single RL agent and require cooperation with other entities. Cooperative multi-agent RL (MARL) research studies the tasks wherein agents need to learn near-optimal policies to maximize a shared team reward while interacting with a stochastic environment and each other.

2.3.1 Multi-Agent RL

The jump from single agent to multi-agent RL presents three challenges. First, there is typically an exponential increase in the state space. This is present in the

ACMP particularly. Indeed, models struggle with the curse of dimensionality. Second, the non-stationary assumption is violated, and the agents' concurrent yet heterogeneous behaviors create an unpredictable environment, even more so than the environment for the single agent. The non-stationarity of the environment stems from the actions of one agent that can yield different rewards depending on the actions of other agents. This becomes particularly troublesome when developing MARL solution approaches.

In model-based single agent RL, the problem is characterized with an MDP. When $n > 1$ agents are involved, a common approach is to extend the MDP model by formulating a stochastic game (SG). This brings in the notion of game theory and is particularly important when the system is competitive, i.e., one agent is performing actions against another agent who is performing counter actions. Therefore, this is not particularly relevant to this thesis, which is focused on cooperative game theory.

However, the canonical structure does provide a framework to assist in formulating multi-agent MDPs. Assuming n agents are in an environment, we can characterize the problem according to a 5-tuple: $(S, U, T, R_{i,\ldots,n}, \gamma)$. $S = S_0 \times S_1 \times \cdots \times S_n$ is the state space, which can be composed of a global state, S_0, that is common to all agents, and local states, S_i, specific to each agent i. $U = A_1 \times \cdots \times A_n$ is the joint action space combining all local action spaces for each agent, where the actions of agents can be fully visible or hidden to other agents depending on problem type. $T : S \times U \times S \to [0, 1]$ is the state transition function and depends on the joint action space and not local action spaces. R_i is the reward function for agent i, which is now dependent on the state space and the joint action space. Finally, γ is the standard discount factor.

Canese et al. (2021) note some additional methodologies to characterize a multi-agent scenario within certain limitations. A partially observable Markov decision

process (POMDP) is another generalization of an MDP that considers the state uncertain. This model is helpful within the limits of a scenario where one agent is being controlled while taking into account the information of another agent that is uncertain or uncontrollable. Some 1v1 WVR ACMP approaches leverage this model. A decentralized partial-observable Markov decision process (Dec-POMDP) extends the POMDP to the classical multi-agent case. A Dec-POMDP gives every agent an observation each time step. Each agent then takes their observation and acts on their own. A policy for a Dec-POMDP maps local (to each agent) observations to actions, which is called the local policy of an agent. These approaches are popular and successful in many applications but still struggle from the curse of dimensionality. An approach that directly tackles this is most likely needed.

In single-agent RL, hierarchical RL is used to combat the curse of dimensionality, and multi-agent RL is already well documented on its challenges with curse of dimensionality. Logically, Makar et al. (2001) model a multi-agent system using a multi-agent hierarchical RL approach. The author notes the use of hierarchy speeds up learning in the multi-agent domain because it focuses on communication at the sub-task level instead of the entire action space. These cooperative sub-tasks are where the agents communicate and are defined at the highest level of the hierarchy. Makar et al. (2001) also explore the cost of communication. This provides a framework to quantitatively report on whether communication is valuable in the 2v1 WVR ACMP, a key focus of this paper. This approach is shown to effectively handle the curse of dimensionality. However, the tradeoff is the large increase in algorithmic complexity and inability to accurately characterize various applications. Before a specific type of SG is selected, it is wise to analyze a few considerations on how the desired multi-agent system operates to determine which is best.

The first consideration is choosing between a centralized and decentralized ar-

chitecture. A centralized architecture models the n agents jointly and solves for a centralized policy for all agents. This leads to a large input and output space. The input space will be the state space in a traditional Markov game or the concatenation of observations in the Dec-POMDP. The output space is the combination of all actions for each agent in each state. It is easy to see how this input increases linearly with the number of agents while the output increases exponentially.

The second option is a decentralized architecture, where agents are modeled and trained independently of each other. This leads to a local policy network for each agent and improves upon the scalability issue relative to the centralized architecture. However, the decentralized architecture violates the stationarity of the environment. Environmental non-stationarity is best characterized in a game of Rock-Paper-Scissors. Two agents are simultaneously performing an action to choose rock, paper, or scissors and then receive some instantaneous reward. However, the reward earned for one agent depends on the action of the other. Therefore, when solving for a policy, the underlying MG would need some notion of probability of each agent picking an action, or some interaction history. If Agent 1 learns a policy to beat Agent 2 quickly, then Agent 2 would react by changing its preference to beat Agent 1, which would lead to changing its policy accordingly. This back and forth continues forever, leading to non-stationarity. While the above example is for a non-cooperative MG, a similar phenomenon occurs in cooperative MGs.

Multi-agent systems can further be characterized along two axes: coordination and communication. Coordination is more closely tied with the type of architecture outlined above and focuses on whether there is a single controlling agent over multiple agents or just multiple agents. Communication focuses on the amount of shared information between agents in the form of state space, previous actions, and so on. Tan (1993) is one of the first to explore the notion of independent versus cooperative

agents. The notion of cooperation is realized through the sharing of instantaneous information, episodic experience, and learned knowledge. The author defines the sharing of instantaneous information through sharing state space, actions, and rewards between agents. In this instance, the information sharing outperforms the independent agents by reward when the information is relevant for the agents. Episodic experience is defined as the communication of a sequence of state, action, and reward triples experienced by the agents. This adds additional information upon the sharing of instantaneous information by bringing in the dimension of time. Learning with episodic experience leads to the same performance as attained by independent learning with independent agents in terms of reward, but it learns the intelligent behavior faster. Tan (1993) finds that the cooperative agents outperform the independent agents if the cooperation is defined and modeled intelligently. The insights and methods of communication provide valuable insight into how AUCAVs may communicate with each other in the MvN WVR ACMP.

2.3.2 Multi-Agent Air Combat

The smallest number of aircraft performing modern air combat missions is the *two-ship section*, commonly just called the *section*. Traditionally, the two aircraft within the section perform two distinct roles: lead and wingman. The lead is responsible for the tactical decisions of the section while the wingman is responsible for covering and supporting the lead. This relationship requires careful coordination of actions; therefore, the modeling approach for the 2v1 WVR ACMP must intelligently characterize the sharing of information.

Shaw (1985) emphasizes the basis of section combat is *mutual support*. There is no explicit definition of mutual support. However, Toubman (2020) infers two concepts that characterize mutual support. The first concept is the creation of *situational*

awareness. It is defined as the gathering of critical information to then be used to inform decisions. For example, an aircraft flying alone has many blind spots, particularly below and behind the aircraft. This limits the ability of the pilot to detect other aircraft visually. In section, the blind spots are significantly reduced if flown correctly. Today, the radar has largely replaced the need to visually detect an aircraft. However, Toubman (2020) notes the same concept still applies. Two radars can cover a larger area than one. An increase in information about the surrounding air of the section informs the execution of better decisions by the section.

The second concept is a *flexible division of roles*. The traditional division of roles is the following. The lead engages the enemy combatant in a 1v1 WVR engagement while the wingman flies above to observe the situation. This arrangement allows the wingman to be in a better position to make tactical decisions for the section. A flexible division of roles allows for the wingman to make that tactical decision by taking over leadership and attacking the opponent that is pursuing the lead. The result is the section's ability to execute offensive and defensive actions simultaneously, which is the reason a two-ship is the smallest unit flown in modern combat.

Determining how the two axes of collaboration within systems, communication and coordination, fit in section combat is critical. Toubman (2020) provides an example of how to merge the basis of section combat with the two axes. Tan (1993) argues the importance of only sharing relevant information being shared between agents. In section combat, relevant information is any information that increases mutual support. Alternatively, we can define relevant information as any information that builds situational awareness and enables a flexible division of roles in section combat. The two axes of multi-agent systems, communication and coordination, are crossed to generate a quadrant, visualized in Figure 6.

TACIT is a decentralized coordination method without communication. In this

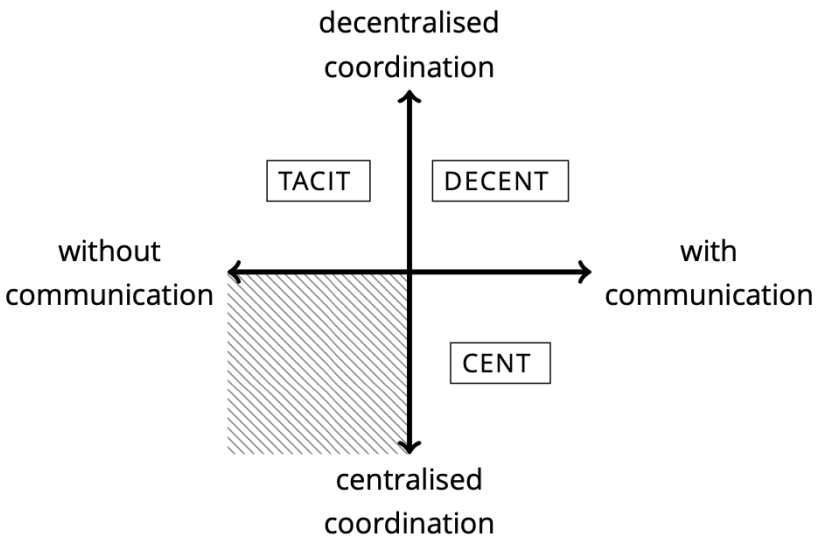

Figure 6. Axes of Team Coordination (Toubman, 2020)

method, the two Blue AUCAVs select their own actions in an individual manner. No inter-AUCAV exchange of information occurs. AUCAVs simply observe each other. Then, each Blue AUCAV performs an action based on these observations. This approach is not consistent with the current operation of section combat and will likely least resemble human fighter pilot tactics. However, this could also be a strength as more effective, yet previously unseen techniques could be developed that are not considered by the human pilots.

CENT is a centralized coordination method with communication. One of the two Blue AUCAVs is designated the lead and becomes the central agent of the section, which means the lead selects the actions for itself and the wingman. The benefit of the wingman is that it sends observations back to the lead to increase the situational awareness of the section. This approach requires a large state space and a large amount of information to be shared repeatedly. However, CENT should theoretically outperform TACIT based on the current human pilot knowledge as the increase in information should lead to higher quality policies and performance.

DECENT is a decentralized method with communication. The two Blue AUCAVs send and receive information between each other, improving the situational awareness of both. Then, the two Blue AUCAVs select their own actions. This provides a more flexible division of roles relative to CENT as there is no single AUCAV designated as the lead, or central agent.

Interestingly, Toubman (2020) finds that the CENT method outperformed the other two methods by generating the most effective behavior and learning the quickest. The author also notes that this insight should not be generalized to every scenario because improvements in the modeling and solving of decentralized methods may improve in the years to come.

Toubman (2020) also contributes the notion that transfer learning can save the training time for agents between ACMPs that are similar. *Transfer learning* is the notion of using knowledge gained on performing one task to the learning of a similar one. This is considered to be one of the leading solutions for improving training times for many practical RL applications. The author demonstrates that transfer learning is appropriate in the ACMP by transferring the knowledge of two agents trained in a 2v1 WVR ACMP to a 2v2 WVR ACMP. The two agents were able to learn the 2v2 WVR ACMP much faster and provided much more effective behavior initially.

The ability to perform transfer learning from the 1v1 WVR ACMP to the 2v1 WVR ACMP may be possible. Ideally, the knowledge transferred would be the ability to maneuver effectively into a position of advantage or out of a position of disadvantage respectively. Then, the remaining knowledge of how to interact and communication between AUCAVs would be learned. There are a few challenges to this, mainly being the notion that a 1v1 engagement is very different than a 2v1 engagement and may transfer useless knowledge. It is not the purpose of this paper to explore this opportunity. However, it provides possible insight into future approaches to moving

from the single-agent to multi-agent scenario.

2.4 Solution Approaches

The continuous and high-dimensional nature of air combat render exact solution approaches found in MDP and DP computationally intractable. Therefore, recent literature has explored the use of techniques found in ADP and RL to elicit high quality behavior from these autonomous aircraft. Each field presents different perspectives on the notion of automating air combat and the specific algorithmic structures proven to be successful.

2.4.1 Approximate Dynamic Programming

The operations research community solves a sequential decision problem from a perspective of making decisions based on the approximate downstream value of being in some state, or ADP (Powell, 2011). In air combat, the initial work approximated this value using a simple least squares regression equation and sound basis function development (McGrew et al., 2010; Fang et al., 2016). While successful, the linearity present in least squares regression appears to place a ceiling on performance, which is logical when you consider the non-linear aspect of an air to air engagement. Wang et al. (2020) test the limits of least squares regression by implementing a 6-DOF nonlinear AUCAV model and non-linear dynamic inverse (NDI) control model. The ADP generated policy slightly outperforms the benchmark despite the non-linear models implemented. Other approaches approximate the value function using neural networks and demonstrate the increased solution quality from the non-linearity functions of neural networks (Jenkins et al., 2021; Crumpacker, 2021). However, a neural network approximation requires more computational effort. The above solution approaches are typically referred to as value function approximations (VFAs) (Powell,

2011).

As research into the ACMP moves into MvN engagements, the field of ADP, and operations research, will be increasingly important as common applications within these fields are fleet management and resource allocation. These application areas contain high-dimensional state spaces, vectored actions (i.e., actions for multiple entities), and the notion of sparse rewards. All three of these characteristics are present in the MvN air combat problem.

2.4.2 Reinforcement Learning

The computer science community solves a sequential decision problem from the notion of an agent(s) that learns by trial and error over time by interacting within an environment, or reinforcement learning (RL). The field of RL and operations research solve a similar type of problem and in a similar manner while using slightly different jargon. For example, RL calls a VFA a critic-only approach. Ma et al. (2018) approach air combat with a critic-only approach, where Deep Q-learning is applied to elicit high quality behavior in a two dimensional fight.

Recent research in RL, including RL applied to air combat, centers around the actor-critic strategy. The *critic* is similar to ADP's notion of approximating the downstream value of states, where RL will expand and sometimes look at the value of state-action pairs. The *actor* is accomplishing a similar goal by trying to create a function that maps from an observation (state) directly to an action. Actor-only approaches are called policy function approximations (PFAs) in the operations research community. The popularity of actor-critics continue to rise for recent successes in AlphaGo and DARPA Dogfight Trials (Pope et al., 2021).

Although the actor critic strategy has proven successful in games such as Atari and air combat simulators, there is an extremely high computational effort required to

implement these algorithms. This computational effort stems from the use of at least two neural networks that must be trained while also accounting for the computational effort to gather information from the model, which can be done through Monte Carlo simulations or smaller n-step trajectories. This *information gather process* is typically resource intensive. Even then, a simple actor-critic with two networks has proven to struggle, particularly where the critic can over-estimate the value of future states. Researchers now add a second critic network and assume the value of the next state is the minimum of the two approximations given by the critic networks.

More advanced algorithms such as the Soft Actor Critic (SAC) will have additional neural networks to adjust and learn specific algorithm hyper-parameters, such as the temperature parameter in SAC (Haarnoja et al., 2018). To date, the SAC algorithm is the superior solution approach in air combat and is proven successful in engagements versus trained military fighter pilots in simulations (Pope et al., 2021). While successful, the exponential computation of resources required to train and adequately explore the hyper-parameters for this research were not possible. Instead, this research leverages a critic-only approach where the value function is approximated using a neural network.

III. Methodology

3.1 Problem Description

In 1v1 air combat, an aircraft is simultaneously performing two tasks: an offensive and defensive mission. The former is where the aircraft is maneuvering into a position of advantage to employ its weapon of choice. The latter is where the aircraft is also maneuvering to avoid positions of disadvantage where the enemy aircraft can employ its weapons of choice. At a high level, in 2v1 air combat, the section of aircraft must still be performing the offensive and defensive missions. However, the section must also communicate dynamically which aircraft is flying the lead position given the current state of the engagement and when to change positions. As we expand into MvN engagements, the information and decisions increase exponentially.

Consider a team of M Blue AUCAVs that identify a team of N Red AUCAVs while out on patrol. The task for the team of M Blue AUCAVs is to determine the sequence of maneuvers as a team that maximize its ability to shoot down all N Red AUCAVs and win the engagement. Ideally, we would simultaneously solve a policy for the Red team; however, resource constraints did not allow this. Therefore, we assume the Red AUCAV utilizes the position-only policy, which was found to be a high quality policy that generates intelligent behavior in the 1v1 WVR ACMP (Crumpacker, 2021). The position-only policy typically incentivizes actions that will drive the Red AUCAV aft of one Blue AUCAV by some range. However, in its current form, the position-only policy is designed for the 1v1 case. Chapter IV modifies the policy for the MvN case. The combination of superior dynamics and high-performing policies by the Red AUCAV should necessitate the novel maneuvering and proper teamwork for the Blue AUCAV(s) in the 1v1 and 2v1 engagement respectively.

There are three key assumptions for the team of Blue AUCAVs. First, we assume

each AUCAV begins the engagement with a full fuel tank. This allows enough time for high-quality engagements to be simulated.

Second, we assume each Blue AUCAV has its own weapons platform: either guns only or guns and a directed energy weapon (DEW). When a Blue AUCAV is equipped with guns, they can be employed when the target is in a weapon engagement zone (WEZ) that starts 500 feet away from the AUCAV and extends 3,000 feet in front within 30 degrees of the center-line. The same constraints apply to the Red AUCAV's gun WEZ.

Third, we assume the communication scheme where all information is processed and acted upon in a central location, i.e., the CENT approach described by Toubman (2020). If other communication forms are desired, a new MDP model would need to be formulated. The section senses and acts as one agent. We assume a shared state space and action space between the section where it is essentially one agent interacting with the environment.

Each Blue AUCAV can end the engagement in the following conditions: killed, crashed, or survived. It is assumed that crashed implies the specific AUCAV ran out of fuel or flew above or below the ceiling and floor the engagement, respectively. In the 2v1 WVR instance, this implies there are 6 possible terminating conditions: (1) the entire section is killed; (2) one AUCAV is killed and the other crashes; (3) one AUCAV is killed and the other survives; (4) the entire section crashes; (5) one AUCAV survives and the other crashes, or (6) the entire section survives.

The final key assumption is perfect information. The team of M Blue AUCAVs knows the position, altitude, and angles of the Red AUCAVs exactly, which prevents the MvN WVR ACMP from needing to be modeled as a state-adversarial MDP (SA-MDP). Also, this assumption is not far from reality based on the recent advancements in technology with semi-conductors, computer vision, and many other technologies.

3.2 MDP Formulation

The Markov decision process (MDP) model creates an architecture for a team of M Blue AUCAVs to learn high-quality policies in a sample efficient manner. The ideal goal for the MDP model is to leverage just enough human insight to elicit emergent behavior that mirrors the tactics and knowledge learned over years of the evolution of fighter combat while also allowing enough freedom for the team to execute unseen tactics that could inform future TTPs. A general MDP formulation is presented that can easily be expanded to the multiple AUCAVs versus multiple AUCAVs context. First, the model utilizes discrete time, defined according to a fixed time step δ^t. The set of decision epochs is given by

$$\mathcal{T} = \{0, \delta^t, 2\delta^t, 3\delta^t, \dots\}. \tag{6}$$

3.2.1 Agent space

Let $\mathcal{Q}$ be the set of AUCAVs in the engagement given by

$$\mathcal{Q} = \mathcal{Q}^B \cup \mathcal{Q}^R, \tag{7}$$

where $\mathcal{Q}^B$ is the set of Blue AUCAVs and $\mathcal{Q}^R$ is the set of Red AUCAVs. No information communication parameters need to be define due to the assumption of perfect information and central processing and communication between the Blue AUCAVs.

3.2.2 State space

We extend Crumpacker (2021) when constructing the state space by expanding to the multi-agent context, providing the Red AUCAV with the 5-DOF kinematic

model, and adding a directed energy weapon (DEW) capability. Let

$$S_t = \left(B_{tq}, R_{tq'}\right)_{q \in \mathcal{Q}^B,\, q' \in \mathcal{Q}^R} \in \mathcal{S} \tag{8}$$

represent the state of the WVR engagement at epoch $t \in \mathcal{T}$ wherein B_{tq} is the status of Blue AUCAV $q \in \mathcal{Q}^B$ and $R_{tq'}$ is the status of Red AUCAV $q' \in \mathcal{Q}^R$ at epoch t.

The state of a Blue AUCAV q is given by the tuple

$$B_{tq} = \left(K_{tq}, M_{tq}\right), \tag{9}$$

wherein K_{tq} and M_{tq} are the statuses of the kinematics and munitions for Blue AUCAV q at epoch t, respectively.

The kinematic status of an AUCAV at epoch t is given by the tuple

$$K_{tq} = \left(x_t, y_t, z_t, V_t, \gamma_t, \chi_t, \alpha_t, \mu_t, T_t^{thtl}, m_t, f_t\right), \tag{10}$$

wherein x_t, y_t, z_t is the coordinate position of the AUCAV in three dimensions in the fixed Earth frame, and V_t is the velocity of the AUCAV along the x axis of the wind reference frame. State variables γ_t and χ_t represent the flight path and heading angles respectively in the fixed Earth frame, and α_t and μ_t are the angle of attack and roll angle of the aircraft respectively. State variables T_t^{thtl}, m_t, and f_t respectively indicate the throttle setting, mass, and amount of fuel remaining at epoch t for the AUCAV.

The munitions status of an AUCAV q and epoch t is given by the tuple

$$M_{tq} = \left(G_{tq}, E_{tq}\right), \tag{11}$$

wherein G_{tq} and E_{tq} are the gun and DEW weapon status at epoch t. G_{tq} is defined

by

$$G_{tq} = (a_{tq}), \tag{12}$$

wherein a_{tq} is the amount of gun ammunition remaining at epoch t for blue AUCAV q. E_{tq} is defined by

$$E_{tq} = (b_{tq}), \tag{13}$$

wherein b_{tq} is the battery level at epoch t for the DEW on blue AUCAV q.

The state of a red AUCAV q' is given by the tuple

$$R_{tq'} = (K_{tq'}, A_{tq'}), \tag{14}$$

wherein $K_{tq'}$ and $A_{tq'}$ are the kinematic status and energy accumulated on red AUCAV q' at epoch t, respectively. This kinematic status is given by the tuple

$$K_{tq'} = \left(x_t, y_t, z_t, V_t, \gamma_t, \chi_t, \alpha_t, \mu_t, T_t^{thtl}, m_t\right) \tag{15}$$

where all variables in this tuple relate to the same definition as present in the Blue AUCAV's kinematic status, K_{tq}, with the fuel variable excluded as we are not focused on controlling when the red AUCAV fires. $A_{tq'}$ is used to keep track of the energy accumulated on the red AUCAV by the DEWs present on the blue AUCAVs.

3.2.3 Action space

The decision for each Blue AUCAV q with decision authority at epoch t is

$$x_{tq}^B = \left(x_{tq}^{\alpha}, x_{tq}^{\mu}, x_{tq}^{thtl}, x_{tq}^G, x_{tq}^E\right),\ \forall\, q \in \mathcal{Q}^B, \tag{16}$$

where x_{tq}^{α} represents the change in angle of attack; x_{tq}^{μ} the change in roll angle; x_{tq}^{thtl} is the change in throttle setting; x_{tq}^{G} is the decision whether to fire the gun or not, and x_{tq}^{E} is the decision whether to fire the DEW or not. If there are AUCAVs with no decision authority, i.e., $\mid \mathcal{Q}^{B-} \mid > 0$, then the same decision will be made at each epoch t by its controller. Let x_t^B represent the decisions for each blue AUCAV at epoch t given by

$$x_t^B = \left(x_{tq}^B\right)_{q \in \mathcal{Q}^B}. \tag{17}$$

Similar to Crumpacker (2021), the sets of allowable state dependent actions for the decision variables governing angle of attack and roll angle are given by

$$\mathcal{X}_q^{\alpha}(S_t) = \left\{ {x_t}_q^{\alpha} \in \{-\delta^{\alpha}, \tfrac{-\delta^{\alpha}}{2}, \tfrac{-\delta^{\alpha}}{5}, 0, \tfrac{\delta^{\alpha}}{5}, \tfrac{\delta^{\alpha}}{2}, \delta^{\alpha}\} : \alpha^L \leq \alpha_t + x_t^{\alpha} \leq \alpha^U \right\}, \tag{18}$$

and

$$\mathcal{X}^{\mu}(S_t) = \left\{ -\delta^{\mu}, \tfrac{-\delta^{\mu}}{2}, \tfrac{-\delta^{\mu}}{5}, 0, \tfrac{\delta^{\mu}}{5}, \tfrac{\delta^{\mu}}{2}, \delta^{\mu} \right\}, \tag{19}$$

wherein δ^{α} and δ^{μ} represent discrete amounts by which we increase or decrease the angle of attack and roll angle respectively. As the above equations show, the preset discrete proportions by which we can change these angles is $1, \frac{1}{2}, \frac{1}{5}, 0$. Equation (18) has the additional parameters α^L and α^U, which represent the maximum angle and minimum angle of attack allowable by the specific air frame. Next, we discretize the throttle setting into discrete sections, given by

$$\mathcal{X}_q^{thtl}(S_t) = \left\{0, 0.2, 0.4, 0.6, 0.8, 1\right\}. \tag{20}$$

Equation (20) indicates that the AUCAV can set the throttle position to some level

between 0 and 1, representing the percent of throttle.

Let $\mathcal{Z}^G_{q,q'}$ be the set of states in which a Blue AUCAV q has a Red AUCAV q' within the gun's weapon engagement zone (WEZ), given by

$$\mathcal{Z}^G_{q,q'} = \left\{S_t : \lambda_{tqq'} < 30^\circ, 500 \leq R(B_{tq}, R_{tq'}) \leq 3000\right\}, \forall\, q \in\ \mathcal{Q}^B,\, q' \in \mathcal{Q}^R \tag{21}$$

where $\lambda_{tqq'}$ and $R(B_{tq}, R_{tq'})$ are the radar angle and range between Blue AUCAV q and Red AUCAV q' at epoch t, respectively. Then, let $\mathcal{Z}^E_{q,q'}$ be the set of states where a Blue AUCAV q is in the DEW's WEZ, given by

$$\mathcal{Z}^E_{q,q'} = \left\{S_t : 45^\circ < \lambda_{tqq'} < 135^\circ, 1000 \leq R(B_{tq}, R_{tq'}) \leq 7000\right\}, \forall\, q \in \mathcal{Q}^B,\, q' \in \mathcal{Q}^R \tag{22}$$

Equation (22) indicates that a DEW can be employed when a Blue AUCAV q is within 1000 and 7000 feet of Red AUCAV q' while being within the side WEZ defined by the radar angle being between 45° and 145°. We can now fully characterize the decision spaces for firing the gun or employing the DEW given by

$$\mathcal{X}^G_q(S_t) = \begin{cases} \{0,1\} & \text{if } a_{tq} > 0,\ S_t \in \mathcal{Z}^G_{q,q'} \\ \{0\} & \text{otherwise} \end{cases}, \tag{23}$$

and

$$\mathcal{X}^E_q(S_t) = \begin{cases} \{0,1\} & \text{if } b_{tq} > l^{\min},\ S_t \in \mathcal{Z}^E_{q,q'} \\ \{0\} & \text{otherwise} \end{cases}, \tag{24}$$

wherein $l^{\min}$ is the minimum battery power required to continually implement the laser. Equation (23) represents when a Blue AUCAV q can fire a gun where 1 rep-

resents firing the gun and 0 represents not firing. A Blue AUCAV q can only fire at Red AUCAV q' when it has ammunition and is within the gun WEZ defined by $\mathcal{Z}^G_{q,q'}$. Equation (24) represents when Blue AUCAV q can employ a DEW where 1 represents continuously applying a laser to the target for the specified time step and 0 represents not employing. Blue AUCAV q can only fire its laser at Red AUCAV q' when there is enough battery energy available, and it is within the DEW WEZ specified by $\mathcal{Z}^E_{q,q'}$.

Let $\mathcal{J} = \{\alpha, \mu, thtl, g, l\}$ represent the index set of decision variables in the decision vector, x^B_{tq}, and the decision spaces. Then, let L^G_{tq} be the G load on AUCAV q at epoch t where L^G_{max} is the maximum allowable G load on the AUCAV. Therefore, the entire set of feasible decisions for a Blue AUCAV q at epoch t is given by

$$\mathcal{X}^B_q(S_t) = \left\{(x^B_{tqj})_{j\in\mathcal{J}} : x_{tqj} \in \mathcal{X}^j_q(S_t),\, L^G_{tq}(S_t, x_{tq}) \leq L^G_{max}\right\}. \tag{25}$$

We have defined the local action spaces for a Blue AUCAV q in $\mathcal{Q}^B$. The joint action space of the multi-agent system can be characterized by

$$\mathcal{X}^B(S_t) = \prod_{q\in\mathcal{Q}^B} \mathcal{X}^B_q(S_t), \tag{26}$$

which indicates the joint action space for all Blue AUCAVs is the Cartesian product of all the local action spaces for each Blue AUCAV q.

Let X^{π^q} represent the decision function based on policy π^q that maps the state space to the action space for the q-th aircraft, i.e., $X^{\pi^q} : \mathcal{S} \to \mathcal{X}^q(S_t),\, S_t \in \mathcal{S}, \forall\, q \in \mathcal{Q}$.

3.2.4 Transitions

There are three possible outcomes of each engagement for each AUCAV: survived, killed, or crashed. There is a distinction between killed and crashed where the for-

mer is when an enemy AUCAV shoots down that AUCAV while the latter is when an AUCAV flies below the floor (10,000 ft) or above the ceiling (50,000 ft) of the engagement. This leads to the three terminal states: Δ_S, Δ_K, and Δ_C, which correspond to survived, killed, and crashed.

The maneuvering and weapon employment decisions for all AUCAVs in the engagement are deterministic under the assumptions of perfect control and information. The stochastic nature of the engagement comes from the outcome of weapon employments. For example, if a blue AUCAV chooses to fire a gun, the result may be the destruction of the red AUCAV based on some probability of kill, p_k. Let S^M denote the system model where the resulting system is updated according to

$$S^M\big(S_t, x_t^B, W_{t+1}\big) = S_{t+1}. \tag{27}$$

The next state of the engagement is determined using the previous state of the engagement, the actions of the blue AUCAVs, and the exogenous information present in the outcomes of weapon employments.

The outcomes of weapon employments are determined by a probability model that takes into account range, radar angle, and aspect angle between two aircraft. First, the probability model scales the range and radar angle information to each be within the range $[0, 1]$. The more desirable radar angles are given values closer to 0 because a radar angle of 0° is ideal. The more desirable ranges are given values closer to 0 as well since being closer to the target is always preferred. To characterize the joint probability space, we subtract each value from 1 and take the product of the two to determine an estimate of the probability of kill.

In explorations through simulation, it was found that the AUCAVs were taking shots facing each other. While this is not uncommon, it is not ideal and is typically a lower probability shot. Therefore, when the aspect angle was greater than 60°,

we subtracted ρ_{kill} from the probability calculated above to ensure more shots were being taken behind aircraft. This led to more realistic simulations. It should be noted that future work should work on formalizing a probability model that is realistic and accurate to modern air combat.

The state of the system at epoch $t+1$ is given by

$$S_{t+1} = \left(B_{t+1,q}, R_{t+1,q'}\right)_{q \in \mathcal{Q}^B,\, q' \in \mathcal{Q}^R} \tag{28}$$

determined via a status update of each AUCAV in the engagement as a part of the system model transition. The status of each blue AUCAV q in the engagement is updated according to

$$B_{t+1,q} = \left(K_{t+1,q}, M_{t+1,q}\right)_{q \in \mathcal{Q}^B}. \tag{29}$$

Let K_B^M denote the kinematic model such that

$$K_B^M(S_t, x_{tq}) = K_{t+1,q} \tag{30}$$

wherein the next state of a Blue AUCAV q at epoch $t+1$ is governed by the previous state of the engagement and its set of actions at epoch t. The kinematic model leverages the 5DOF kinematic model as described by Crumpacker (2021). The individual updates of each variable within the tuple K_{tq} occur as follows:

$$x_{t+1,q} = x_{tq} + \left(V_{t+1,q} \cos\gamma_{t+1,q} \cos\chi_{t+1,q}\right)\delta^t \tag{31}$$

$$y_{t+1,q} = y_{tq} + \left(V_{t+1,q} \cos\gamma_{t+1,q} \sin\chi_{t+1,q}\right)\delta^t \tag{32}$$

$$z_{t+1,q} = z_{tq} + \left(V_{t+1,q} \sin\gamma_{t+1,q}\right)\delta^t \tag{33}$$

$$V_{t+1,q} = V_{tq} + \frac{\delta^t}{m_{tq}}\left(T\cos\alpha_{t+1,q} - D - W\sin\gamma_{tq}\right) \tag{34}$$

$$\gamma_{t+1,q} = \arctan\left(\frac{\frac{\delta^t}{m_{tq}}(T\cos\alpha_{t+1,q} + L\cos\mu_{t+1,q} - W\sin\gamma_{tq})}{V_{t+1,q}}\right) \tag{35}$$

$$\chi_{t+1,q} = \arctan\left(\frac{\frac{\delta^t}{m_{tq}}(L\sin\mu_{t+1,q})}{V_{t+1,q}}\right) \tag{36}$$

$$\alpha_{t+1,q} = \alpha_{tq} + x^{\alpha}_{tq} \tag{37}$$

$$\mu_{t+1,q} = \mu_{tq} + x^{\mu}_{tq} \tag{38}$$

$$T^{thtl}_{t+1,q} = x^{thtl}_{tq} \tag{39}$$

$$m_{t+1,q} = m_{tq} - \left(x^{g}_{tq} r^{g} m^{g}\right)\delta^t \tag{40}$$

$$f_{t+1,q} = f_{tq} - \left(cT\right)\delta^t \tag{41}$$

Let G^M represent the gun transition model of a Blue AUCAV q, where

$$G^M(S_t, x^B_{tq}) = G_{t+1,q}, \tag{42}$$

and where $G_{t+1,q}$ is updated based on the decision whether or not the gun is fired for Blue AUCAV q. This can be computed by

$$a_{t+1,q} = \begin{cases} a_{tq} - r^g\delta^t & \text{if } x^g_{tq} = 1 \\ a_{tq} & \text{otherwise ,} \end{cases} \tag{43}$$

where r^g is the rate of fire for the gun. Similarly, let E^M represent the DEW transition model of a Blue AUCAV q, where

$$E^M(S_t, x^B_{tq}) = E_{t+1,q}, \tag{44}$$

and where $E_{t+1,q} = b_{t+1,q}$ is updated based on the status of the DEW at epoch t and

the action taken. This battery update can be computed by

$$b_{t+1,q} = \begin{cases} b_{tq} - r_d^b \delta^t & \text{if } x_{tq}^l = 1 \\ \min\{b_{tq} + r_c^b \delta^t, 1\} & \text{if } x_{tq}^l = 0,\ b_{tq} < 1 \\ b_{tq} & \text{otherwise} \end{cases}, \tag{45}$$

where r_d^b and r_c^b are the rate at which the DEW battery decays when employed and charges when not in use and not at full battery, respectively. The update $B_{t+1,q}$ is now fully characterized.

The status of each red AUCAV q' in the engagement is updated according to

$$R_{t+1,q'} = \left(K_{t+1,q'}, A_{t+1,q'}\right)_{q' \in \mathcal{Q}^R}. \tag{46}$$

The kinematic status $K_{t+1,q'}$ is updated according to the kinematic model K_R^M. If we were to solve for a policy for the Red AUCAV(s) as well, this problem would become a Markov game and leverage game theory. Therefore, due to time constraints and complexity, this thesis assumes a pre-determined policy to control the Red AUCAV(s). Hence

$$K_R^M(S_t, \pi_{q'}^R) = K_{t+1,q'} \tag{47}$$

where $\pi_{q'}^R$ is the *position-only* policy, which determines decisions for a Red AUCAV q'. It is assumed that a Red AUCAV's decisions are reactions to the Blue AUCAV(s) decisions. Therefore, the tuple $K_{t+1,q'}$ is computed using the same set of equations as $K_{t+1,q}$.

The position-only policy, $\pi_{position}^R$, leverages two reward shaping techniques found in the air combat maneuvering problem (ACMP) literature. First, a range score is

calculated that scales all the range information between Red AUCAV q' and a Blue AUCAV q into the interval $(0, 1]$ according to

$$S^R_{q',q} = e^{\frac{-|R(B_{tq},R_{tq'})-R_d|}{\kappa_r}}, \forall\, q' \in \mathcal{Q}^R,\, q \in \mathcal{Q}^B, \tag{48}$$

where R_d and κ_r are tunable parameters that control the desired range and decay of the range score as range increases. Crumpacker (2021) identifies $R_d = 1,000$ and $\kappa_r = 2,000$ as well-tuned values. The second reward shaping function is the angle score and scales all the angle information to $[0, 1]$. The score for Red AUCAV q' is given by

$$S^A_{q',q} = 1 - \frac{\lambda_{tq'q} + \epsilon_{tq'q}}{360}, \forall\, q' \in \mathcal{Q}^R,\, q \in \mathcal{Q}^B. \tag{49}$$

Overall, a Red AUCAV q' will choose an action according to

$$\underset{x_{tq'} \in \mathcal{X}^{\mathcal{R}}(S_t)}{\operatorname{argmax}} \left\{ \max_{q \in \mathcal{Q}^B} \left\{ S^R_{q',q} S^A_{q',q} \right\} \right\}, \tag{50}$$

which indicates the Red AUCAV chooses actions that maximize its position advantage for the Blue AUCAV q with the maximum range and angle score. This can be thought of as a "greedy" policy where the red AUCAV engages the Blue AUCAV has the highest position-only score relative to the Red AUCAV's current position at epoch t.

Finally, let $A^M_{q'}(S_t, x^B_{tq})$ denote the energy accumulation model where

$$A^M_{q'}(S_t, x^B_{tq}) = A_{t+1,q'}, \forall\, q' \in \mathcal{Q}^R. \tag{51}$$

The energy accumulated on Red AUCAV q' at the next step is based on the previous state of the system and the actions taken by the various Blue AUCAVs in the

engagement. The energy accumulation is updated according to

$$A_{t+1,q'} = \begin{cases} A_{tq'} + \iota \sum_{q \in \mathcal{Q}^B} \mathbb{I}(x_{tq}^B = 1) S\big(R(B_{tq}, R_{tq'}), \lambda_{tq}\big) & \text{if } \exists\, q \in \mathcal{Q}^B \text{ s.t. } x_{tq}^E = 1 \\ A_{tq'} - \kappa & \text{if } x_{tq}^E = 0 \,\forall\, q \in \mathcal{Q}^B,\ A_{tq'} > 0 \\ A_{tq'} & \text{otherwise} \end{cases} \tag{52}$$

where ι is the fraction of energy absorbed by the Red AUCAV based on its material, $S\big(R(B_{tq}, R_{tq'}), \lambda_{tq}\big)$ is the intensity of the beam as a function of the distance and the radar angle between the Blue AUCAV q and the Red AUCAV q', and κ is the rate at which energy dissipates from the material's surface. Therefore, the first condition indicates that for all Blue AUCAVs currently employing their DEW against a Red AUCAV, the energy accumulates according to the fraction of energy that makes it through the material. The second condition indicates the decay of energy from the material when no DEW is employed. Finally, the third condition indicates the resting energy accumulated on a Red AUCAV is 0.

3.2.5 Contributions

The salient contributions are earned when the system evolves to one of the terminal states where one of the AUCAVs in the engagement is killed. Otherwise, the system earns no contribution. Note, we condense the outcomes into a set $\mathcal{O} = \{S, K, C\}$ and Δ_o^q means that AUCAV q is in terminal state $o \in \mathcal{O}$. The contributions are formally

defined as

$$C\big(S_t, x_t, S_{t+1}\big) = \begin{cases} \sum_{q\in\mathcal{Q}^B} \mathbb{I}(S_t \notin \Delta, S_{t+1} = \Delta_S^q)\xi & \text{, if } S_t \notin \{\Delta_o\}_{o\in\mathcal{O}} \\ \sum_{q'\in\mathcal{Q}^R} \mathbb{I}(S_t \notin \Delta, S_{t+1} = \Delta_S^{q'})(-\xi) & \text{, if } S_t \notin \{\Delta_o\}_{o\in\mathcal{O}} \\ \sum_{q\in\mathcal{Q}^B} \mathbb{I}(S_t \notin \Delta, S_{t+1} = \Delta_C^q)(-\nu) & \text{, if } S_t \notin \{\Delta_o\}_{o\in\mathcal{O}} \\ \sum_{q'\in\mathcal{Q}^R} \mathbb{I}(S_t \notin \Delta, S_{t+1} = \Delta_C^{q'})\nu & \text{, if } S_t \notin \{\Delta_o\}_{o\in\mathcal{O}} \\ 0 & \text{, if } S_t \notin \{\Delta_o\}_{o\in\mathcal{O}},\ S_t \notin \{\Delta_o\}_{o\in\mathcal{O}} \\ 0 & \text{, otherwise} \end{cases} \tag{53}$$

where ξ is a large positive contribution, and ν is a smaller (relative to ξ), yet also large positive contribution where $\nu \ll \xi$. The first two conditions present in Equation (53) indicate the system will incur a large reward or penalty when an enemy or friendly AUCAV is killed, respectively. The third and fourth conditions indicate a moderate contribution will be incurred when one AUCAV crashes or is killed before shooting down the Red AUCAV, respectively. The fifth condition is explicitly defined to highlight that no contribution is earned when a the system transitions from one non-terminal state to another non-terminal state. This is different from previous iterations of the ACMP. In initial training, the system earned a -1 for each non-terminal transition; however, this led to issues with the AUCAV learning to kill itself quicker and receive a slightly higher contribution than if it executed a long engagement and then was shot down. No extra contribution is earned when one AUCAV is shot down while the other shoots down the Red AUCAV.

3.2.6 Objective function and Bellman equation

The objective of this thesis is to obtain an optimal policy, π^*, for the Blue AUCAVs that maximizes the expected total discounted contributions (ETDC). In the

centralized case, a single policy must be obtained, $\pi^* = \pi^B$, for the ETDC equation given by

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi} \Big[\sum_{t=1}^{\infty} \gamma^{t-1} C(S_t, X^{\pi}(S_t), S_{t+1}) \Big], \tag{54}$$

where $\gamma \in [0, 1)$ is the discount factor. Recall a single policy is determined because of the assumption of a joint action space between the Blue AUCAV(s). Let $x_t = X^{\pi}(S_t)$ represent the policy, or decision function, that returns the decision, x_t, given the state of the engagement, S_t. To get to this policy, we must determine a solution to the Bellman Equation given by

$$J(S_t) = \max_{x_t^B \in \mathcal{X}(S_t)} \mathbb{E}\Big[C(S_t, x_t^B, S_{t+1}) + \gamma J(S_{t+1}) \mid S_t, x_t^B \Big] \tag{55}$$

Unfortunately, obtaining an optimal policy is computationally intractable with exact approaches found in classical dynamic programming. Therefore, we implement an ADP algorithm to obtain high-quality policies that can defeat the Red AUCAV.

3.3 Model-based Reinforcement Learning Approach

In this research, we employ an approximate policy iteration (API) strategy to solve for a high-quality maneuver and shooting policy for the Blue AUCAV(s) by approximating the value function. To increase the rate of learning, basis functions are leveraged to capture information on the important features of the engagement. This solution approach pulls aspects from the API strategies found in Jenkins et al. (2021) and Crumpacker (2021) by approximating the value function based on neural network learning around the pre-decision state. However, this approach considers more than one hidden layer in the neural network architecture and tests for the best activation function instead of predetermining which one to use.

3.3.1 Basis Functions

Let $\mathbf{\Phi}(S_t)$ be the basis function vector containing the set of features. This research pulls from the existing literature to create a representative vector that can accurately characterize the state of the engagement and assist in learning high-quality maneuvers. A majority of these basis functions are derived from Wang et al. (2020) while the few others are derived from Crumpacker (2021) due to these use of a 5-DOF kinematic model. In this research, for the 1v1 Guns Only problem instance, we defined the basis function as

$$\mathbf{\Phi}(S_t) = \Big[\lambda_{tB},\, \epsilon_{tB},\, S_A^B,\, S_R^B,\, \chi_{tB} - \chi_{tR},\, \gamma_{tB} - \gamma_{tR},\, V_{tB} - V_{tR}, \tag{56}$$

$$x_{tB} - x_{tR},\, y_{tB} - y_{tR},\, z_{tB} - z_{tR}, a_{tB}^{\text{standardized}},\, PE^B,\, KE^B\Big] \tag{57}$$

where $a_{tB}^{\text{standardized}} = \frac{a_{tB}}{a_{0B}}$ is the ammunition of blue at epoch t, and PE^B and KE^B respectively refer to the potential and kinetic energy of the aircraft at epoch t. Potential energy and kinetic energy of the Blue AUCAV at epoch t are given by

$$PE^B = (m_{tB} + f_{tB}) g z_{tB}, \tag{58}$$

$$KE^B = \frac{1}{2}(m_{tB} + f_{tB}) V_{tB}^2. \tag{59}$$

Similar to McGrew et al. (2010) and Crumpacker (2021), this research explores first order interactions and second order effects. This research additionally explores third order effects. After initial testing, all interaction terms were removed from the basis function due to time constraints and lower quality results. Finally, we scale all elements of $\mathbf{\Phi}(S_t)$ to the range $\big[0, 1\big]$ to reduce the effects of different scales across the various basis functions.

3.3.2 Exploration Mechanism

A key tradeoff in any RL or ADP solution approach is exploration versus exploitation. In a value function approximation approach, there are two standard methods: selecting actions via ϵ-greedy policy or exploring starts. In the ϵ-greedy policy, we select an action at random with probability ϵ while selecting the action that maximizes our current estimate of our value function with probability $1 - \epsilon$. In exploring starts, we randomly start in a new portion of the state space each episode or training step. This research employs a variant of the exploring starts mechanism, where the random starts will continually adapt to assist in combating the sparse rewards problem.

3.3.3 Sparse Reward Mechanism

The existing literature in the ACMP consistently faces the sparse reward problem, commonly called the credit assignment problem. The only areas of high contribution are when AUCAVs kill, get killed, or crash. In this research, everywhere else in the state space will typically transition to earn a contribution of 0. Therefore, some mechanism is needed to disperse rewards from these fewer areas of higher contribution to the much larger area of little contribution. The common approach is to create reward shaping functions, where we modify the small negative contribution we typically earn in some manner to encourage the agent to move toward those areas of higher contribution. This mechanism induces human bias and fails to leverage an important contribution of ADP/RL research, where the agents learn and exhibit new behaviors that can inform new tactics in air combat.

This research will leverage the adaptive sampling mechanism formulated by Crumpacker (2021), modifying it slightly. In essence, we break down the state space into high-quality and low-quality samples. We define high-quality samples as randomly generated states that result in aircraft starting in each others' WEZ, i.e., states where the

areas of high contribution can easily be determined. We define low-quality samples as the rest of the state space that is outside the various WEZ regions. Our goal is to decrease the number of high-quality samples visited during algorithm execution as we learn more and disperse the observed higher contribution throughout the rest of our state space. We modify this mechanism by splitting results to be in offensive and defensive positions, where Blue starts behind or in front of Red respectively when facing the same heading.

The numbers of high-quality and low-quality samples depend on the episode number, k, and a tunable parameter, β_p, which controls the rate of decay for our high-quality samples. Formally, we calculate the number of high-quality and low-quality samples, N_k^H and N_k^L using

$$N_k^H = \lceil p_k N \rceil, \tag{60}$$

$$N_k^L = N - N_k^H, \tag{61}$$

where p_k is the percentage of high-quality points needed and is governed by

$$p_k = \frac{1}{k^{\beta_p}}. \tag{62}$$

Figure 7 displays how different values of β_p impact the number of high-quality samples within the algorithm across $k = 20$ episodes.

This research employs the set of Red AUCAVs with an ability to shoot. Therefore, as previously mentioned, we split the number of high-quality and low-quality points into offensive and defensive scenarios according to some probability. In initial testing, this probability is set to be around 0.55 as we want to allow the agent more opportunities to learn when to shoot.

Figures 8-10 shows the adaptive state sampling scheme in relation to the three

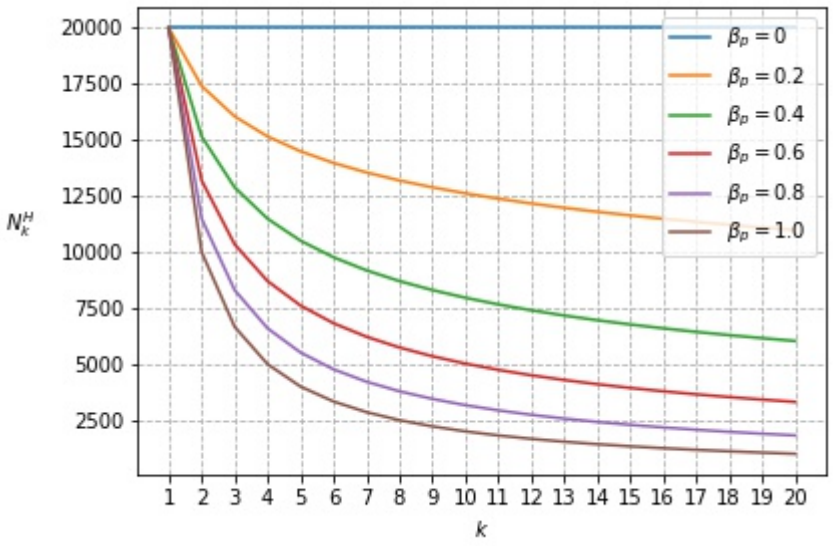

Figure 7. Effect of β_p on N_k^H where $L = 20,000$

dimensional coordinates of the Blue AUCAV assuming $\beta_p = 0.6$. The green points represent high quality samples, N_k^H, while the grey points represent low quality samplers, N_k^L. Note only high quality samples are considered in the first episode while only mainly low quality samples are considered in the final episode, $k = 10$. The intuition is the value learned in the first few episodes, where most samples are high quality, is dissipated through out the rest of the state space in such a manner that the low quality samples at episode 10 do not suffer from sparse rewards.

3.3.4 Algorithm Formulation

This research employs an n-layer feed-forward neural network, where n is the number of layers and n_H is the number of hidden layers in the network. The relationship of n and n_H is given by

$$n = n_H + 2 \tag{63}$$

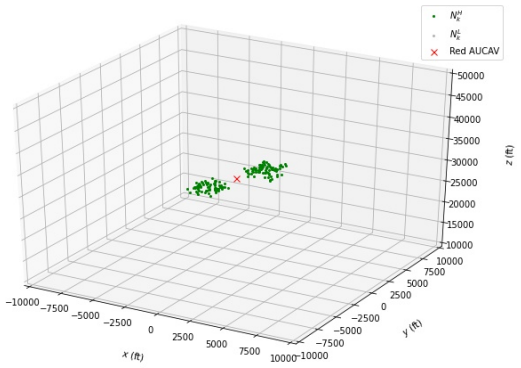

Figure 8. Adaptive State Sampling Scheme, $k = 1$

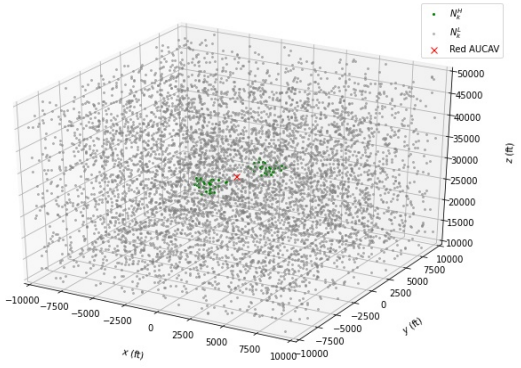

Figure 9. Adaptive State Sampling Scheme, $k = 3$

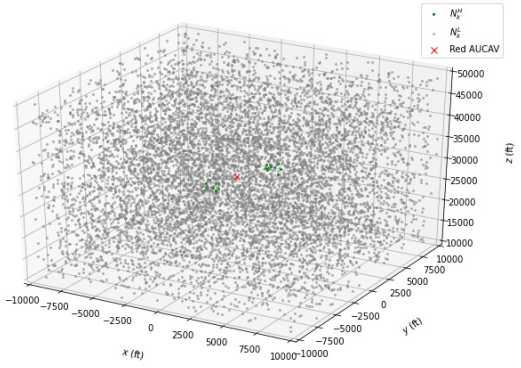

Figure 10. Adaptive State Sampling Scheme, $k = 10$

where the architecture includes an input layer, n_H hidden layers, and an output layer. Each hidden layer i consists of a set of activation units, $\mathcal{H}^i = \{1, 2, \ldots, |\mathcal{H}^i|\}$, $i = 1, \ldots, n_H$. The input layer produces $|\mathcal{H}^1|$ outputs, which are given by

$$\Upsilon_h^{(2)} = \sum_{f \in \mathcal{F}} \Theta_{f,h}^{(1)} \phi_f(S_t), \ \forall\, h \in \mathcal{H}^1, \tag{64}$$

wherein $\Theta^{(1)} \equiv \left[\Theta_{f,h}^{(1)}\right]_{f \in \mathcal{F}, h \in \mathcal{H}^1}$ is an $|\mathcal{F}| \times |\mathcal{H}^1|$ matrix that controls the mapping from the input layer to the hidden layer. This is only for the input layer to the first hidden layer. Multiple activation functions are tested including the leaky rectified linear unit (ReLU), parameterized ReLU, sigmoid, hyperbolic tangent, and ReLU. After initial testing, it is apparent that the ReLU activation function is giving the highest quality results. The standard ReLU activation function is given by

$$\text{ReLU}(z) = \max\left\{0, z\right\}. \tag{65}$$

ReLU is applied at each perceptron to provide inputs into the hidden layer, given by

$$Z_h^{(2)} = \text{ReLU}\left(\Upsilon_h^{(2)}\right), \ \forall\, h \in \mathcal{H}^1. \tag{66}$$

If there is more than one hidden layer, then the hidden layer i produces $|\mathcal{H}^{i+1}|$ outputs, which are given by

$$\Upsilon_h^{(i+1)} = \sum_{f \in \mathcal{F}} \Theta_{f,h}^{(i)} Z_h^{(i)}, \ \forall\, h \in \mathcal{H}^i, \ i = 1, \ldots, n_H \tag{67}$$

wherein $\Theta^{(i)} \equiv \left[\Theta_{f,h}^{(i)}\right]_{f \in \mathcal{F}, h \in \mathcal{H}^i}$, and the linear transform is repeatedly applied based on the number of hidden layers. Then, the ReLU activation is applied at each perceptron

within each of these hidden layers, given by

$$Z_h^{(i+1)} = \text{ReLU}\big(\Upsilon_h^{(i+1)}\big),\ \forall\, h \in \mathcal{H}^i,\ i = 1, \ldots, n_H \tag{68}$$

The final hidden layer, n_H, produces a single output, given by

$$\Upsilon_h^{(n_H+2)} = \sum_{h \in \mathcal{H}^{n_H+1}} \Theta_h^{(n_H+1)} Z_h^{(n_H+1)} \tag{69}$$

wherein $\Theta_h^{(n_H+1)} \equiv \big[\Theta_H^{(n_H+1)}\big]_{h \in \mathcal{H}^{n_H+1}}$ is an $|\mathcal{H}^{n_H}| \times 1$ matrix of weights to map from the final hidden layer to the output layer. The output layer does not apply an activation function, given by

$$\bar{J}(S_t \mid \Theta) = \Upsilon_h^{(n_H+2)} \tag{70}$$

wherein $\Theta = (\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(n_H+2)})$ is the dense tuple of neural network weights. This research implements the Mean Squared Error (MSE) loss function, or the squared L2 norm, alongside the adaptive moment estimation (ADAM) optimizer in Pytorch 1.7.1.

The algorithm now makes decisions leveraging the policy

$$X_{\text{ADP}}^{\pi^B}(S_t \mid \Theta) = \underset{x_t^B \in \mathcal{X}^B(S_t)}{\text{argmax}} \Big\{ \mathbb{E}[C(S_t, x_t^B, S_{t+1})] + \gamma \bar{J}(S_{t+1} \mid \Theta) \Big\} \tag{71}$$

This leads to the following state value function

$$\bar{J}(S_t \mid \Theta) = \mathbb{E}[C(S_t, X_{\text{ADP}}^{\pi^B}(S_t \mid \Theta)), S_{t+1}] + \gamma \mathbb{E}[\bar{J}(S_{t+1}|\Theta)|S_t, x_t^B] \tag{72}$$

The neural network approximated state value function is now fully characterized. Algorithm 1 presents the formalized algorithm used in this research. This algorithm

is a modified variant of the algorithms presented by Crumpacker (2021) and Jenkins et al. (2021) with more in-depth exploration of the tunable parameters such as number of hidden layers and activation functions.

Algorithm 1. API-NN to determine π^B

Initialize Θ^0.
for $k = 1, \ldots, K$ **do**
    Generate L random samples of states.
    for $l = 1, \ldots, L$ **do**
        Transform to basis function space, $\mathbf{\Phi}(S_t)$.
        Predict and record the predicted value using Θ^{k-1}.
        Record the realized value using Equation (73).
    end for
    Compute the new weights, $\hat{\Theta}^k$, using predicted and realized values.
    Update Θ^k using Equation (74).
end for
return decision function $X^{\pi^B}_{ADP}(\cdot|\Theta^K)$ for policy π^B.

To start, the weights of the neural network are initialized using the Xavier Uniform initialization, which is one of the many techniques to initialize weights to small, random values near zero. (Mishkin and Matas, 2015; Glorot and Bengio, 2010). Then, we start training for K episodes. At the start of each episode, we generate a random sample of states using the adaptive state sampling scheme detailed in Section 3.3.3. Once this set is generated, the algorithm visits each of these sampled states, transforms the state to the basis function space, and then computes a realized value of the current policy at episode k using

$$\hat{j}^l_t = \max_{x^B_t \in \mathcal{X}^B(S_t)} \mathbb{E}[C(S_t, x^B_t, S_{t+1})] + \gamma \bar{J}(S_{t+1}|\Theta^k). \tag{73}$$

After visiting all L sampled states during episode k, the algorithm executes a policy improvement step before the end of the episode. The sample neural network weights, $\hat{\Theta}^k$, are computed using the Adam optimizer. Then, our algorithm updates

the neural network weight tuple according to

$$\Theta^k = \alpha_k \hat{\Theta}^k + (1 - \alpha_k)\Theta^{k-1} \tag{74}$$

wherein $\alpha_k \in [0, 1]$ is the smoothing rule parameter. In this thesis, the polynomial smoothing rule is leveraged. The polynomial learning rate is given by

$$\alpha_k = \frac{1}{k^{\beta_\alpha}} \tag{75}$$

wherein β_α is a tunable parameter.

IV. Testing, Results, and Analysis

In this chapter, we introduce a representative scenario of the air combat maneuvering problem (ACMP) in 1v1 and 2v1 cases across three different problem instances to analyze the MvN ACMP developed in Chapter III. The efficacy of the ADP solution is examined as we quantitatively compare the ADP-generated policies against the current benchmark policy. The robustness of the ADP solution, with regards to hyper-parameters and problem type, is tested using computational experiments, which also function as a tuning mechanism for the ADP solution to maximize performance. All computational experiments are conducted using an Intel(R) Xeon(R) Silver 4210, 2.20GHz, 10-Core processor with 128GB of RAM in Python 3.6 using Numpy Version 1.19.5 and PyTorch 1.7.1.

4.1 Representative Scenario Description

This research examines the scenario wherein M Blue autonomous unmanned combat aerial vehicles (AUCAVs) engage with $N = 1$ superior Red AUCAVs. The notion of *superior* refers to the ability of the Red AUCAV to change its angle of attack and roll angle slightly more each simulation step, i.e., superior dynamics. This leads to the Red AUCAV being able to turn tighter radii. The Red AUCAV employs the position-only benchmark policy as discussed in Chapter III. When the engagement starts, the M Blue AUCAVs seek to maneuver and shoot down the superior Red AUCAV. Due to time constraints, a time limit for each engagement is assumed at 180 seconds, or three minutes, of engagement time.

Within our representative scenario, we evaluate performance across three different problem instances, defined via the starting state of the engagement: offensive, defensive, or neutral. To adequately explore each of these problem instances, we gen-

erate 100 starting states for each, first fixing the location of the Red AUCAV and then randomly displacing the Blue AUCAVs within the desired region. Figures 11, 12, and 13 show these displacements in the offensive, defensive, and neutral setting, respectively.

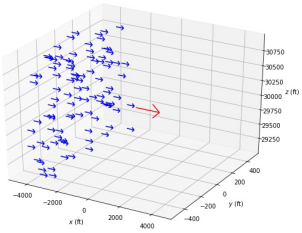

Figure 11. Offensive Starts

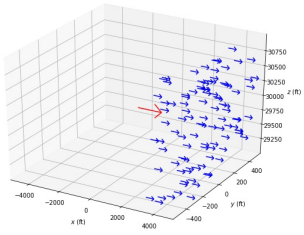

Figure 12. Defensive Starts

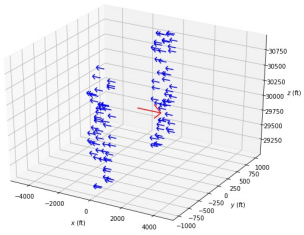

Figure 13. Neutral Starts

4.2 Benchmark Policy

A benchmark policy is established as a function by which we can assess the quality of our ADP solution approach. Ideally, we could compare our ADP generated policy to the optimal policy. However, as previously mentioned, this is impossible due to the continuous and high-dimensional state space. Therefore, a benchmark policy is leveraged. Crumpacker (2021) finds the previously mentioned position-only and position-energy policy to be high-quality. We employ the position-energy policy as our benchmark policy for the Blue AUCAVs. Built on the reward shaping functions developed by Fang et al. (2016) and McGrew et al. (2010), the position-energy policy is comprised of three components, or scores.

The first score is the range score, given by

$$S_R = e^{-\frac{|R(B_t,R_t)-R_d|}{\kappa_r}}, \tag{76}$$

where $R(B_t, R_t)$ is the range between the Blue and Red AUCAV, R_d is the desired

range to be at for the Blue AUCAV, and κ_r is a range decay parameter. The second score is the angle score, given by

$$S_A = 1 - \frac{\lambda_B + \epsilon_B}{360}, \tag{77}$$

where λ_B and ϵ_B are the radar and aspect angles of the Blue AUCAV respectively. Finally, the third score is the energy score, given by

$$S_E = \begin{cases} 1 & \text{if } k > 2 \\ \frac{1}{2} + \frac{k - k^{-1}}{3} & \text{if } \frac{1}{2} \leq k \leq 2 \\ 0 & \text{otherwise} \end{cases}, \tag{78}$$

where $k = \frac{H_{E,B}}{H_{E,R}}$ is the specific energy ratio of the Blue AUCAV to the Red AUCAV where $H_E = H + \frac{V^2}{2g}$, wherein H is the altitude, V is the velocity, and g is the standard acceleration due to gravity. Thus, a Blue AUCAV implementing the benchmark position-energy policy makes decisions according to

$$\underset{x_t \in \mathcal{X}^B(S_t)}{\operatorname{argmax}} \Big\{ \omega S_A S_R + (1 - \omega) S_E \Big\}, \tag{79}$$

where $\omega \in (0, 1)$ is a tunable parameter that is set to 0.97 after initial exploration and testing. In the 2v1 engagement, each Blue AUCAV makes a decision according to this policy independently.

There are two insights to note about this benchmark. First, there is no formulation of firing in the position-energy policy. Therefore, it is assumed that all shots are taken when possible. This already makes the position-energy policy formidable because even if our ADP-generated policy learns to also take shots whenever it can, it will still be taking the same number of shots as the benchmark. Second, this benchmark

policy drives its AUCAV behind its opponent and maintains a higher energy level than its opponent – which is an extremely aggressive, offensive strategy. The main disadvantage of this strategy is its lack of defensive ability.

Once a superlative policy is determined for each problem instance, the superlative policies are simulated for 100 replications and their performances are compared to those attained by this benchmark policy. Then, the behaviors elicited by the superlative policies are cross referenced with Shaw (1985) to determine if there is any emergent behavior.

4.3 1v1 Designed Experiment and Analysis

The purpose of the experimental design is to determine the highest performing combination of hyper-parameters. In this research, performance is specifically a function of solution quality, computational effort, and robustness (Barr et al., 1995). Solution quality is measured by number of wins attained over replications across defensive, offensive, and neutral problem instances. Computational effort is measured by the time it takes to train given a specific hyper-parameter combination. Robustness can use variance across hyper-parameter combinations as a proxy.

There are eight different tunable parameters, or factors, for the ADP solution procedure, which are summarized in Table 1. The various levels for which these factors can be tested were determined through multiple, sequential iterations of fractional designs to slowly zoom in on the hyper-parameter space shown.

A full factorial is then designed to enumerate each possible combination of these hyper-parameters, which leads to a total of 288 design runs. The average probability of kill, $p_{kill,r}$, for replication r is calculated using

$$p_{kill,r} = \sum_{t=0}^{T_{end}} p_{kill} \tag{80}$$

Table 1. 1v1 Factors and Levels

Factor	Description	Levels
n_H	number of hidden layers	{2, 3}
β^{α}	decay of smoothing rule	{0.4, 0.6}
β^{p}	decay of high-quality samples	{0.6, 0.8}
$\mathcal{H}$	number of activation units	{27, 64, 96}
η	regularization parameter (L2)	{0, 0.001}
α_{NN}	learning rate of neural net	{0.01, 0.2, 0.3}
K	number of episodes	{10, 15}
L	number of samples	{10000}

where T_{end} is the time step at which the engagement ends. We execute $r = 1, \ldots, R$ replications and then calculate the average probability of kill for that policy using

$$\bar{p}_{kill} = \frac{1}{R} \sum_{r=1}^{R} p_{kill,\,r}. \tag{81}$$

If an AUCAV does not have shot opportunity at time step t, then p_{kill} for that AUCAV at time step t is 0. This pushes the average, $\bar{p}_{kill\,r}$, downward to zero. However, we believe this downward pressure helps evaluate the quality of policies. Consider a policy that rarely maneuvers behind the Red AUCAV but by chance gets some high quality shot for one time step. When we use Equation (80), this policy is not considered higher quality when compared to a policy that continually has Red within its gun WEZ where there are minimal time steps with $p_{kill} = 0$. This metric rewards policies who maintain advantage against an adversary more than policies who periodically get high probability shots. Table 2 displays the top three combinations for each of the three problem instances and sorts according to Equation (80) across 10 replications versus the superior Red AUCAV in that problem instance.

Table 2 shows that the superlative policy in a offensive, defensive, and neutral problem instance are Runs 79, 87, and 117 respectively. At first glance, it is not intuitive why the Blue AUCAV's $\bar{p}_{kill}$ is highest in the defensive start. In the offensive

Table 2. 1v1 Designed Experiment Top Results

	Blue $\bar{p}_{kill}$	Time *(hrs)*	Run	API-NN Parameters							
				n_H	β^{α}	β^{p}	$\mathcal{H}$	η	α_{NN}	K	L
Offensive	**0.0008**	**2.94**	**79**	**3**	**0.6**	**0.6**	**27**	**0.001**	**0.2**	**10**	**10,000**
	0.0006	4.01	257	3	0.4	0.6	96	0.001	0.01	15	10,000
	0.0005	3.89	286	2	0.6	0.8	96	0.001	0.3	15	10,000
Defensive	**0.0049**	**3.12**	**87**	**3**	**0.6**	**0.8**	**64**	**0.001**	**0.2**	**10**	**10,000**
	0.0043	3.10	63	3	0.6	0.8	64	0	0.2	10	10,000
	0.0039	3.91	111	3	0.6	0.8	64	0	0.3	15	10,000
Neutral	**0.0011**	**3.89**	**117**	**3**	**0.4**	**0.8**	**96**	**0**	**0.3**	**10**	**10,000**
	0.0010	3.15	249	3	0.4	0.6	64	0	0.2	10	10,000
	0.0003	2.69	286	2	0.6	0.8	96	0.001	0.3	15	10,000

start, the Blue AUCAV will be behind and have the Red AUCAV within a 30° radar angle. That satisfies only 2 of the 3 requirements to have a non-zero probability of kill. The p_{kill} equals 0 at the start because the range between the aircraft is over $4,000$ ft. Therefore, if the policy is unable to make the maneuvers to decreasing range while marinating the angle, the Blue AUCAV can easily never see a shot opportunity, which is what happens in the ADP generated policies. This is why the neutral and offensive $\bar{p}_{kill}$ are lower and directly map to the results Table 3. To evaluate, these superlative policies are simulated and compared with the benchmark policy, position-energy, over 100 replications. Table 3 summarizes the results.

Table 3. 1v1 Overall Results

	Number of Blue Wins		
	Offensive	Defensive	Neutral
Benchmark	**63**	22	24
ADP	34	**40**	**25**

Table 3 shows that the ADP generated policies perform poorly in the offensive problem instance while outperforming in the defensive and neutral problem instances in terms of outcome of the engagements. It is important to note that number of blue wins incorporates shooting down the Red AUCAV as well as inducing the Red AUCAV to crash. Sections 4.3.1, 4.3.2, and 4.3.3 respectively explore the offensive,

defensive, and neutral problem instances in detail to determine the performance and quality of the ADP generated policies.

4.3.1 Offensive Starts

When the Blue AUCAV is starting in a position of advantage, or an offensive start, a desirable outcome is a win that takes the least amount of time while maintaining that position of advantage as much as possible. In this research, we count a win as the ability of the Blue AUCAV to shoot and kill the Red AUCAV or the ability to induce the Red AUCAV to crash below the floor of the engagement. This is the main statistic of focus. The secondary and tertiary statistics are the time of the engagement and the ability to maintain advantage. The time of the engagement in seconds can be computed directly from the number of epochs times the fixed time step in the simulation. The ability to maintain a position of advantage and avoid a position of disadvantage can be approximated by calculating the average probability of kill for the Blue and Red AUCAV respectively.

In Section 4.3, Table 2 shows that the superlative policy in the offensive problem instance comes from Run 79. To compare against the benchmark policy, this ADP generated policy, $\pi^{\text{Run 79}}$, is simulated 100 times from offensive starting positions to determine its solution quality. These offensive positions are visualized in Figure 11. Table 4 summarizes these results.

Table 4. 1v1 Offensive Performance over 100 Replications

Policy	Blue Won	Time (s)		Blue $\bar{p}_{kill}$		Red $\bar{p}_{kill}$	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	**63**	30.20s	6.34s	0.0109	0.0080	**0.0013**	**0.0004**
ADP, $\pi^{\text{Run 79}}$	34	29.98s	3.23s	0.0022	0.0012	0.0041	0.0020

Table 4 shows the inability of our ADP solution procedure to learn a policy that outperforms the benchmark. In terms of time, there is not statistical difference be-

tween the ADP and benchmark policy in terms of ending engagements quicker. In terms of maintaining advantage, it is clear that the benchmark policy is higher quality when considering the number of wins but also the confidence interval for the Red AUCAV's probability of kill, which is lower than the confidence interval from $\pi^{\text{Run 79}}$ in a statistically significant manner.

The possible reasons for poor performance relative to the benchmark are numerous. However, the determination of findings do reinforce the high quality of the position-energy policy from previous works, particularly in the offensive problem instance (Crumpacker, 2021). There are two key reasons why the position-energy policy is a formidable policy. First, the policy provides the Blue AUCAV with the sole purpose of maneuvering to some distance R_d behind the Red AUCAV and maintaining a higher energy (through velocity or altitude) than the Red AUCAV. This implies the position-energy policy is aggressive, a characteristic that works well in an offensive setting, as reinforced in Table 4. Second, we assume the position-energy benchmark policy takes all possible shots. This makes it hard for our ADP generated policy because if our ADP generated policy does not learn to shoot whenever possible, it will automatically be at a disadvantage – think shots on goal in sports. The more shots a Blue AUCAV takes, the more likely the Blue AUCAV will win the engagement.

Figures 14, 15, and 16 display a sample offensive engagement wherein the Blue AUCAV wins using the ADP generated policy. Note there are circles with a number inside on the plots. Each number within a circle represents an influential point in time for the simulated engagement and increment chronologically. Influential Point 0 refers to the starting states of the engagement, where the red and blue circle represents the Red AUCAV and Blue AUCAV positions, respectively. The highest numbered circles represent the positions of the AUCAVs when a kill shot occurs and is successful. The closed blue circle represent the Blue AUCAV firing a shot successfully while the red

$\times$ represents the Red AUCAV being shot.

At Influential Point 0, the Red AUCAV starts around $4,000$ feet in front of the Blue AUCAV while both are facing in the positive x-axis direction. As we move forward in time to Influential Point 1, the Red AUCAV makes a wide turn to the right, or negative x-axis direction, while the Blue AUCAV, according to our ADP policy, executes a tight turn to the left while descending in altitude. At Influential Point 1, the Blue AUCAV then increases the slope of its descent while switching its heading to face more in the negative y-axis direction. This forces the Red AUCAV to dive down after the Blue AUCAV, where the Red AUCAV earns a shot opportunity as shown in Figure 17. However, it is unsuccessful and the Blue AUCAV gets a similar shot opportunity a few time steps later and is able to kill the Red AUCAV.

Another method of visualizing the engagement is in terms of the air combat geometry angles, i.e., the radar angle, λ, and the aspect angle, ϵ. When λ_B is closer to 0, this means the angle between the Blue AUCAV's velocity vector and the line of sight vector from the Blue AUCAV to the target is closer to 0, which translates to facing directly at the Red AUCAV. When this value is closer to 180, this translates to facing directly away from the Red AUCAV. On the other hand, when ϵ_B is closer to 0, this means the angle between the line of sight vector from the Blue to the Red AUCAV and the velocity vector of the Red AUCAV are facing similar directions, which translates to the Blue AUCAV being behind the Red. When ϵ_B is closer to 180, the Red AUCAV is behind the Blue.

Figure 18 shows the Blue AUCAV starting in the Offensive quadrant where its radar angle is below $20°$, which indicates it is facing directly at the Red AUCAV, while its aspect angle is below $20°$, which indicates it is behind the Red AUCAV. However, as the engagement continues, the Blue AUCAV, when utilizing the ADP generated policy, quickly loses its advantage and is in a position of disadvantage by

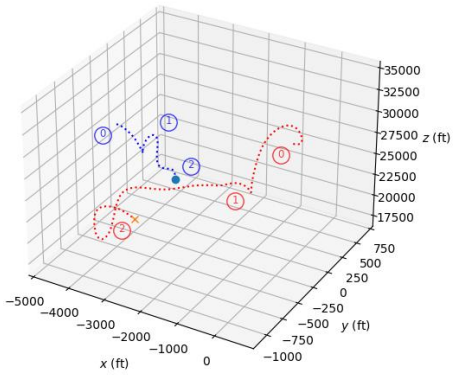

Figure 14. 1v1 **Offensive Engagement in** xyz **Plane**

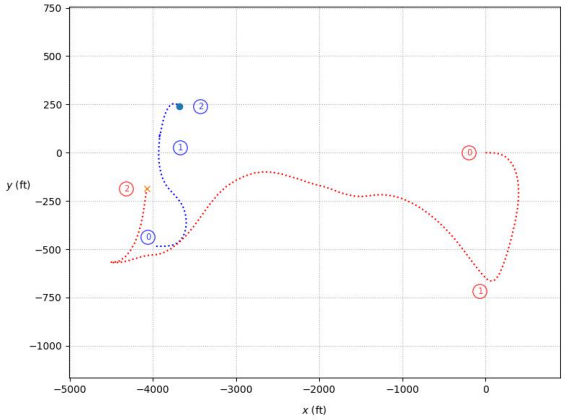

Figure 15. 1v1 **Offensive Engagement in** xy **Plane**

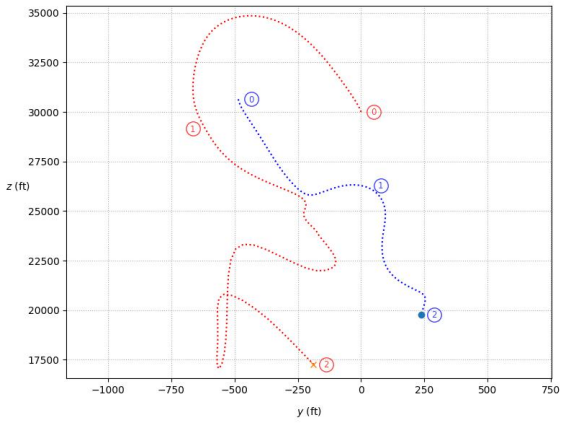

Figure 16. 1v1 **Offensive Engagement in** xz **Plane**

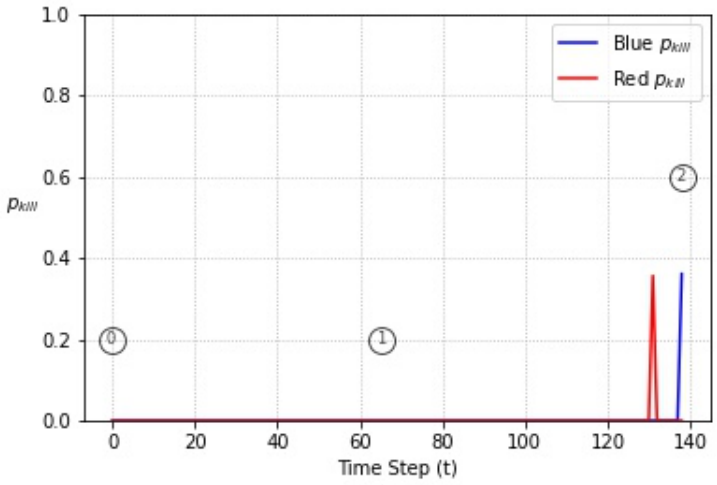

Figure 17. 1v1 **Offensive Engagement:** p_{kill} **Trends**

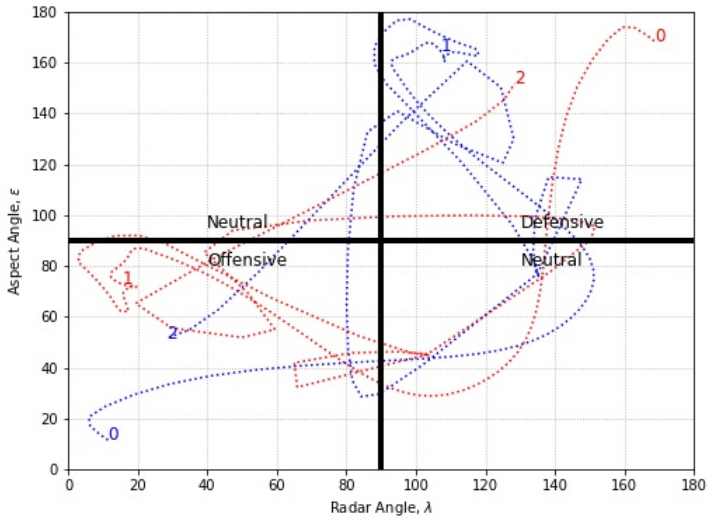

Figure 18. 1v1 **Offensive Engagement: Angle Trade-offs**

Influential Point 1. The inability to maintain that position of advantage is a key reason in why our ADP generated policy fails to outperform the benchmark.

4.3.2 Defensive Starts

When a Blue AUCAV starts from a position of disadvantage, or a defensive start, desirable behavior changes from the offensive start. Now, a desirable behavior is the ability to win. If a win is not possible, then desirable behavior can mean lasting as long as possible to maximize the chances of either a friendly AUCAV coming to assist or the Red AUCAV making some sub-optimal maneuver. In Section 4.3, Table 2 shows that the superlative policy for the defensive problem instance comes from Run 87. To evaluate, we simulate this policy, $\pi^{\text{Run 87}}$, 100 times using the various starting positions visualized in Figure 12. Table 5 summarizes these results.

Table 5. 1v1 Defensive Performance over 100 Replications

Policy	Blue Won	Time (s)		Blue p_{kill}		Red p_{kill}	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	22	31.68s	7.16s	0.0009	0.0003	0.0068	0.0014
ADP, $\pi^{\text{Run 87}}$	**40**	9.71s	1.26s	**0.0050**	**0.0015**	0.0079	0.0019

The ADP generated policy outperforms the benchmark in terms of total number of engagements won over 100 replications. There are two other key insights from Table 5. First, the engagements when using the ADP generated policy are significantly shorter. Second, when using the ADP generated policy, the Blue AUCAV is better at attaining a higher probability of kill over the engagement. These two insights are analyzed further as we examine a sample engagement visualized in Figures 19, 20, and 21.

Each circle within a number represents an influential point in time for the simulated engagement. The blue and red circles with a 0, or Influential Point 0, represent the starting points for the Blue and Red AUCAV, respectively. At the start of the engagement, both aircraft are facing in the positive x-axis direction with the Blue AUCAV starting with a slight altitude advantage and slightly offset in the negative y direction. Figure 20 shows that as the engagement begins, the Blue AUCAV turns

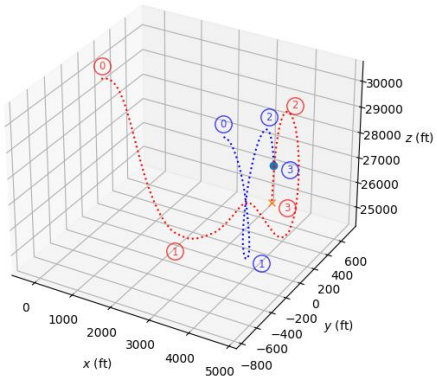

Figure 19. 1v1 **Defensive Engagement in** xyz **Plane**

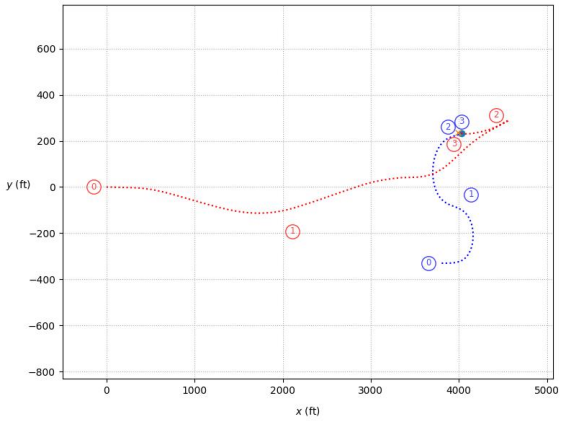

Figure 20. 1v1 **Defensive Engagement in** xy **Plane**

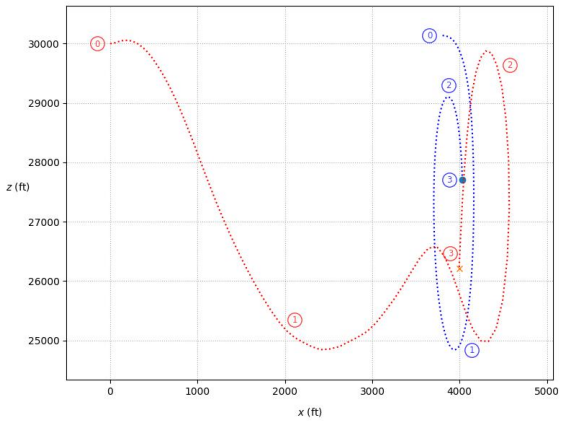

Figure 21. 1v1 **Defensive Engagement in** xz **Plane**

toward the Red AUCAV. This is a consistent occurrence in almost all of the defensive replications using the ADP policy.

At first, this seems illogical. However, it is important to remember the assumptions of the problem. The Red AUCAV is superior in terms of agility and it is starting behind. Combine this with the defensive start and the Blue AUCAV is an precarious position. The position-energy policy is focused on getting to the distance R_d behind the Red AUCAV while still facing directly at it. This leads to taking wider turns and maintaining a similar flight path and heading angle. Meanwhile, the ADP generated policy adopts a high-risk, high-reward approach. Figures 20 and 21 show the Blue AUCAV simultaneously turning toward the Red AUCAV while diving down to gain energy respectively.

At Influential Point 1, the Red AUCAV is continuing to close in on the Blue AUCAV while also diving down and gaining energy. As this is happening, the Blue AUCAV begins to loop back up and gain altitude over the Red AUCAV, which forces the Red AUCAV to follow. It is interesting to note that the Red AUCAV has shot opportunities from the time at Influential Point 1 and Influential Point 2. However, all of these shots are low probability shots that are less than 0.05. As the Blue AUCAV approaches Influential Point 2, it begins to decelerate and slowly turn back down to the ground. Meanwhile, the Red AUCAV continues to increase its velocity and also turn back down to follow the Blue AUCAV. This is the crucial point of the engagement.

As we transition to Influential Point 3, because the Blue AUCAV decelerated and the Red AUCAV accelerated, the Red AUCAV flies past the Blue AUCAV on the downward trajectory and is unable to slow down due to gravity. Now, at Influential Point 3, the Blue AUCAV is able to take the kill shot and win the engagement.

Figure 22 shows the evolution of these angles over the engagement. The numbers

within the graph correspond to the four Influential Points defined earlier. The top left quadrant relates to a neutral scenario as neither AUCAV is really behind the other despite one AUCAV facing at the other. This is where head on shots would reside. The top right quadrant represents a defensive scenario as the enemy AUCAV is behind the AUCAV, given by a high ϵ, and the AUCAV is not facing towards the enemy AUCAV, given by a high λ. Note this is where the Blue AUCAV starts the engagement since we are analyzing a defensive start. The bottom left shows an offensive scenario as an AUCAV is behind and facing relatively at the enemy AUCAV. Note this is where the Red AUCAV starts. Finally, the bottom right quadrant indicates a neutral scenario.

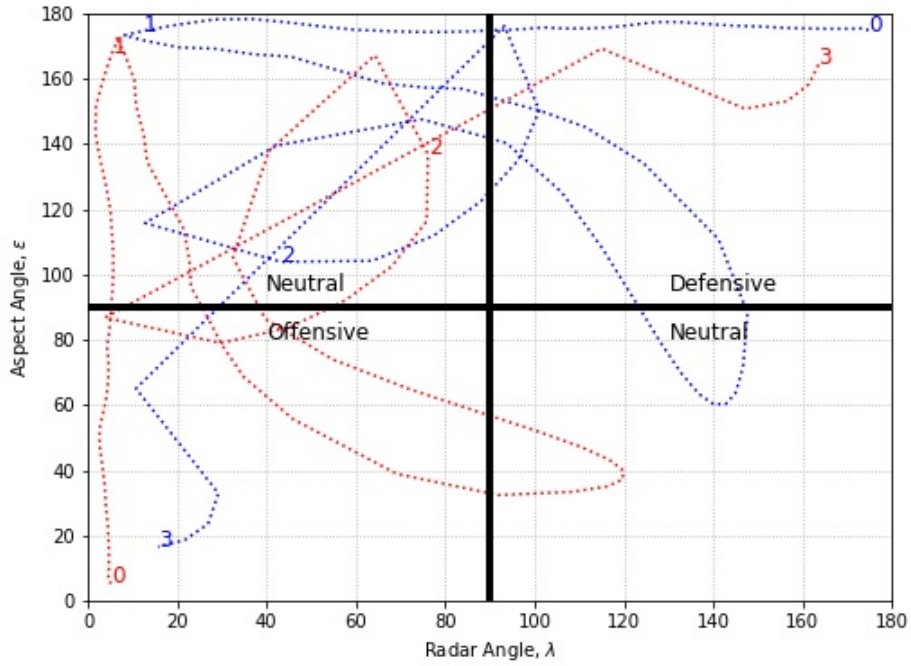

Figure 22. 1v1 **Defensive Engagement: Angle Trade-off**

Figure 22 shows how the Blue AUCAV starts in a defensive scenario, i.e., Influential Point 0, maneuvers over time to gain an angles advantage, and takes the kill shot at Influential Point 3. Conversely, note how the Red AUCAV loses advantage over

time where its angles change in a symmetric manner to the Blue AUCAV's angles. This is because these angles are interrelated with each other. The radar angle for a Blue AUCAV at time t and the aspect angle for the Red AUCAV at time t are supplementary, and add up to 180°. Note at Influential Point 0, the radar angle for the Blue AUCAV is nearly 180° while the aspect angle for the Red AUCAV is nearly 0°. This is why the Blue AUCAV and the Red AUCAV appear to mirror each other in Figure 22. Another insight is that the Blue AUCAV is never truly in a position of advantage until the last second before it takes the kill shot. This is reinforced in Figure 23.

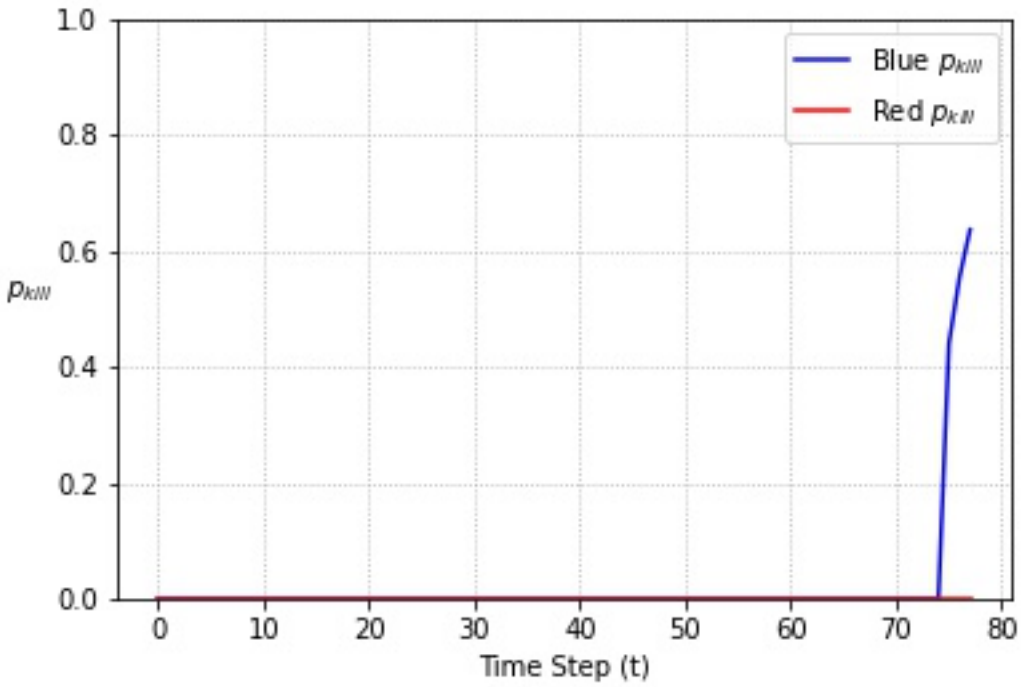

Figure 23. 1v1 **Defensive Engagement:** p_{kill} **Trends**

An interesting behavior occurs even if the Blue AUCAV is unable to shoot down the Red AUCAV at this point of the engagement. Consider the scenario where the Blue AUCAV misses the shot on the Red AUCAV. The Red AUCAV will accelerate down to gain energy once again while the Blue AUCAV accelerates after Red. The

Red AUCAV continues to accelerate away from the Blue AUCAV while trying to pull up. However, it is unable to pull up in time and crashes below the floor of our engagement. This *high risk, high reward* behavior elicited using the ADP policy is the main factor in why the engagements last much less time than when using the benchmark.

4.3.3 Neutral Starts

In a neutral start, initial maneuvers from the starting point will be key factors in whether or not the Blue AUCAV is able to win the engagement. Therefore, after the ability to outperform the benchmark in terms of wins, a desirable behavior is higher average probability of kill for the Blue AUCAV and a lower average probability of kill for the Red AUCAV relative to the benchmark. The time of the engagement is not as important as in the offensive problem instance. Table 2 from Section 4.3 identifies Run 117 as the superlative policy for the neutral problem instance. This superlative policy, $\pi^{\text{Run 117}}$, is executed 100 times using the various starts visualized in Figure 13. Table 6 summarizes these results.

Table 6. 1v1 Neutral Performance over 100 Replications

Policy	Blue Won	Time (s)		Blue $\bar{p}_{kill}$		Red $\bar{p}_{kill}$	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	24	76.69s	11.29s	0.0010	0.0003	0.0036	0.0010
ADP, $\pi^{\text{Run 117}}$	25	61.3s	6.86s	0.0013	0.0005	0.0052	0.0013

Table 6 shows that our ADP generated policy slightly outperforms the benchmark policy by winning 1 more engagement. There are two key insights from this table. First, the length of the engagements when utilizing either policy are relatively the same, with the benchmark policy lasting slightly longer. Second, the ADP generated policy yields a wider confidence interval for the Blue AUCAV's average probability of kill due to the induced policy either gaining advantage or disadvantage quickly.

Figures 24, 25, and 26 display a sample engagement from the neutral start.

Influential Point 0 indicates that the Red AUCAV starts facing the positive x axis direction and the Blue AUCAV starts facing the negative x axis direction. As the engagement begins, both aircraft initiate a turn to the negative y-axis direction, or South, and descend in altitude. Influential Point 1 shows the Red AUCAV turning slightly tighter and tracking toward Blue in the x direction. However, as this is happening, Influential Point 1 also shows that the Blue AUCAV begins a steep ascent away from the Red AUCAV.

As time passes from Influential Point 1 to Influential Point 2, the Blue AUCAV keeps climbing over $20,000$ feet while the Red AUCAV pulls back down and flies aimlessly. This seems illogical at first. However, the reason for this aimless behavior from the Red AUCAV is due to an extremely clever maneuver by the Blue AUCAV. Remember the Red AUCAV makes decisions according to the position-only policy, which is the product of the range and angle scores, $S_R \cdot S_A$. The range score has a tunable parameter, κ_r, that controls the decay of this range score as our range increases away from our desired range R_d, which we set equal to $2,000$. The Blue AUCAV learns to take advantage of this decay parameter and accelerates away from the Red AUCAV, which renders the Red AUCAV paralyzed as any action it can take returns a score of 0.

Figure 27 shows the scores for the actions the Red AUCAV takes according to the position-only policy. Note the steep descent from Influential Point 1 to Influential Point 2 in the Red AUCAV's score due to the Blue AUCAV's steep ascent in altitude. This is the reason for the *aimless behavior* of the Red AUCAV in the sample neutral engagement. While this may not be a realistic tactic, it is a salient result that the ADP generated policy learns this emergent behavior that exploits the mathematics of Red's policy to its advantage. However, this exploitation is worthless if the Red

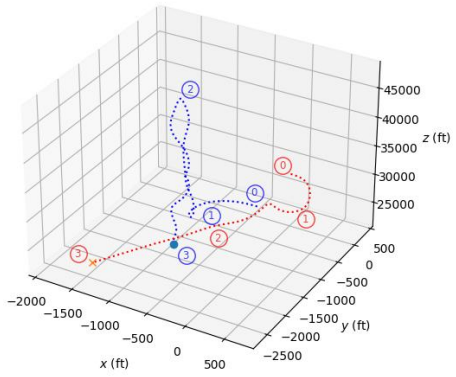

Figure 24. 1v1 **Neutral Engagement in** xyz **Plane**

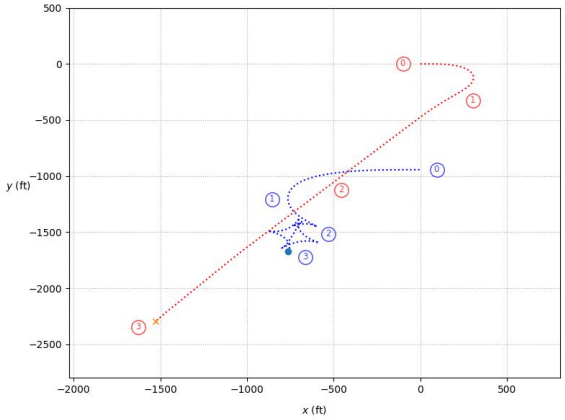

Figure 25. 1v1 **Neutral Engagement in** xy **Plane**

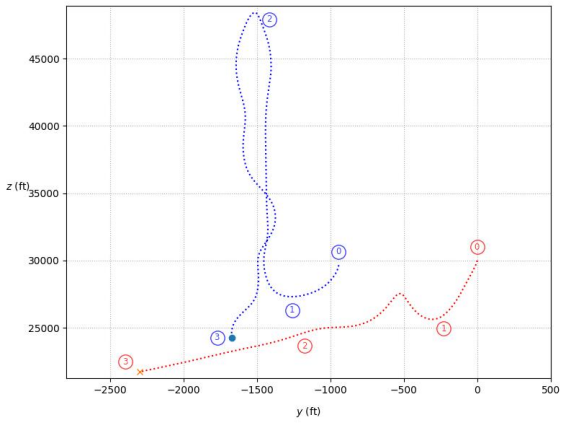

Figure 26. 1v1 **Neutral Engagement in** yz **Plane**

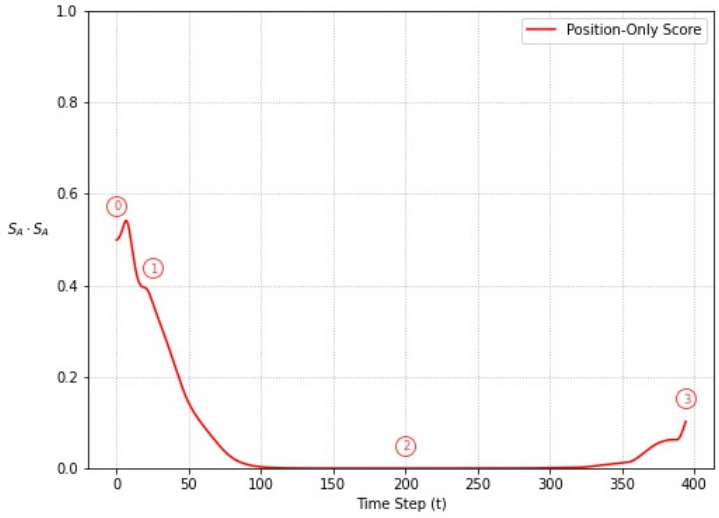

Figure 27. Position-Only Scores for Red AUCAV 1v1 Neutral Engagement

AUCAV is able to pick up a signal and maneuver behind the Blue AUCAV again.

To avoid this, the Blue AUCAV executes a diving maneuver and attacks the Red AUCAV from above at an extremely high velocity. This diving maneuver gains so much velocity that by the time the Red AUCAV begins to pick up a signal and maneuver against the Blue AUCAV, a kill shot is taken and the Blue AUCAV wins the engagement.

Figure 28 shows the angle tradeoff for the engagement as both start exactly at the origin, where both their radar and aspect angles are set to 90°. At the beginning of the engagement, the Red AUCAV takes the advantage first. Note the Red AUCAV almost bringing its radar angle to blue under 30° before Influential Point 1. The Blue AUCAV is able to maneuver away and then execute that steep ascent that induces the Red AUCAV into a sparse reward situation. For the rest of the engagement, the Blue AUCAV slowly works it way back to neutral and then into a position of advantage before it takes the kill shot at a relatively high aspect angle.

While this shot was successful, it is not the highest quality shot in terms of the

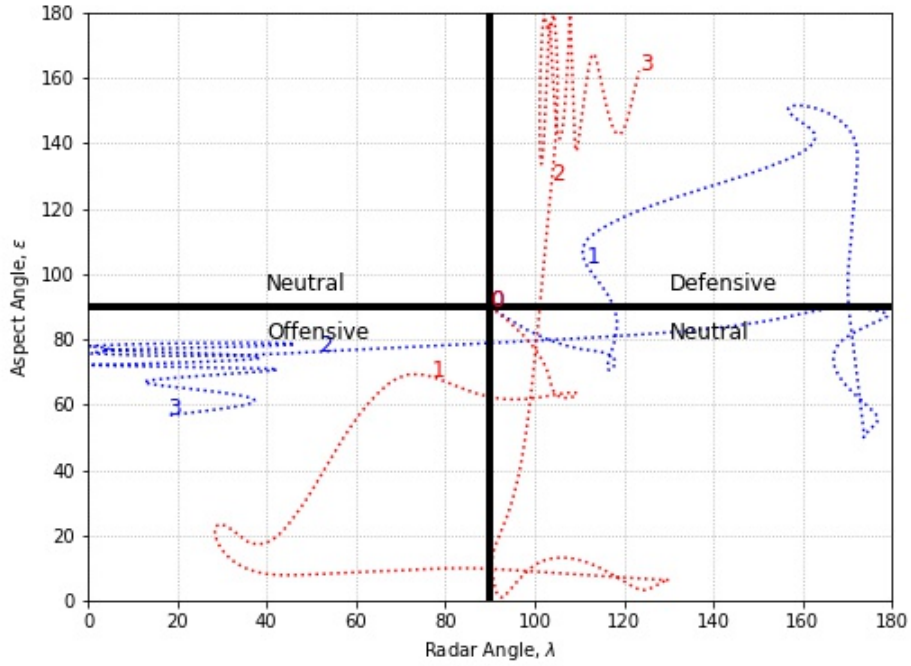

Figure 28. 1v1 **Neutral Engagement: Angle Trade-off**

magnitude of the p_{kill}. The reason for this is the range from which the shot is taken is over $2,000$ ft and at a radar angle of just under $20°$, which are both in the lower half of their desired values. Other replications that mimic this maneuver will wait a few more steps before taking the kill shot with a much higher p_{kill}.

In review, the Blue AUCAV learns to induce the Red AUCAV with a sparse reward situation where every possible action results in a value of 0 according to the position-only policy. Then, the Blue AUCAV executes a high altitude attack and is able to kill the Red AUCAV before it can outmaneuver the Blue AUCAV.

4.4 2v1 Designed Experiment and Analysis

Similar to the 1v1 analysis, a designed experiment is constructed to determine the highest performing combination of hyper-parameters, where performance is a function

of solution quality, computational effort, and robustness (Barr et al., 1995). There are eight tunable parameters, or factors, for the ADP solution procedure, as shown in Table 7.

Table 7. 2v1 Factors and Levels

Factor	Description	Levels
n_H	number of hidden layers	$\{5\}$
β^{α}	decay of smoothing rule	$\{0.4\}$
β^{p}	decay of high quality samples	$\{0.2, 0.6\}$
$\mathcal{H}$	number of activation units	$\{164, 196\}$
η	regularization parameter (L2)	$\{0, 0.001\}$
α_{NN}	learning rate of neural net	$\{0.1, 0.2\}$
K	number of episodes	$\{10\}$
L	number of samples	$\{10000\}$

Multiple layers and activation units are tested in the initial screening and tuning. However, it is apparent that a 5 hidden layer architecture is best suited for the 2v1 instance. Also, the size of the 2v1 designed experiment is much smaller due to resource constraints, particularly when it came to evaluating the resulting policies. For example, in the 1v1 case, a single engagement can be simulated in under five minutes, which lead to a quick evaluation, allowing for more hyper-parameter to be tested. In the 2v1 case, a single engagement could take anywhere from 1 to 2 hours.

The long simulation times are due to the curse of dimensionality with respect to our action space. In the 1v1 instance, our action space has a cardinality of 588. In the 2v1 instance, our action space has a cardinality of $345,744$. Therefore, when comparing hyper-parameter runs within our designed experiment, we only simulate the resulting policy 3 times across offensive, defensive, and neutral problem instances. Then, once a superlative policy is determined, the policy is compared with the benchmark over 20 replications.

We also restrict the action space when training to always allowing the opportunity for an AUCAV to shoot, which reduces the cardinality of our action space to $86,436$.

If a Blue AUCAV decides to shoot despite not being allowed to because it is not within the Gun WEZ, then the p_{kill} for that shot is 0. This reduction allows for quicker training times and saves memory. If we do not use the reduced action space, we store $2,765,952$ bytes for just a simple column vector containing the values of the resulting next states. When we reduce the action space, we only need to store $691,488$ bytes. Note these byte values are based on utilizing Numpy in Python 3.6.

A full factorial experiment is designed to explore the hyper-parameter space for performance. Table 8 displays the hyper-parameter combinations and the resulting performance over 5 replications. Table 8 shows that the superlative policy in an offensive, defensive, and neutral problem instance are Runs 16, 4, and 7, respectively for the 2v1 engagement. To evaluate, these superlative policies are simulated for 20 replications and compared with the benchmark policy, individual position-energy. Table 9 summarizes the overall results.

Table 8. 2v1 Designed Experiment Top Results

Run	Blue 1 $\bar{p}_{kill}$			Blue 2 $\bar{p}_{kill}$			API-NN Parameters							
	Offensive	Defensive	Neutral	Offensive	Defensive	Neutral	n_H	β^α	β^p	$\mathcal{H}$	η	α_{NN}	K	L
1	0.0021	0.0000	0.0127	0.0000	0.0057	0.0000	5	0.4	0.6	164	0	0.1	10	10,000
2	0.0212	0.0000	0.0116	0.0000	0.0056	0.0000	5	0.4	0.2	164	0	0.1	10	10,000
3	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	5	0.4	0.6	196	0	0.1	10	10,000
4	0.0001	**0.0059**	0.0111	0.0000	**0.0041**	0.0000	5	0.4	0.2	196	0	0.1	10	10,000
5	0.0321	0.0000	0.0123	0.0000	0.0058	0.0000	5	0.4	0.6	164	0.001	0.1	10	10,000
6	0.0000	0.0000	0.0084	0.0000	0.0057	0.0000	5	0.4	0.2	164	0.001	0.1	10	10,000
7	0.0000	0.0000	**0.0205**	0.0000	0.0000	0.0000	5	0.4	0.6	196	0.001	0.1	10	10,000
8	0.0021	0.0000	0.0058	0.0000	0.0000	0.0000	5	0.4	0.2	196	0.001	0.1	10	10,000
9	0.0041	0.0000	0.0112	0.0000	0.0057	0.0000	5	0.4	0.6	164	0	0.2	10	10,000
10	0.0074	0.0000	0.0126	0.0000	0.0058	0.0000	5	0.4	0.2	164	0	0.2	10	10,000
11	0.0000	0.0058	0.000	0.0000	0.0000	0.0000	5	0.4	0.6	196	0	0.2	10	10,000
12	0.0000	0.0043	0.0133	0.0000	0.0000	0.0123	5	0.4	0.2	196	0	0.2	10	10,000
13	0.0421	0.0051	0.0007	0.0000	0.0000	0.0000	5	0.4	0.6	164	0.001	0.2	10	10,000
14	0.0032	0.0000	0.0004	0.0000	0.0043	0.0000	5	0.4	0.2	164	0.001	0.2	10	10,000
15	0.0213	0.0000	0.0127	0.0000	0.0059	0.0000	5	0.4	0.6	196	0.001	0.2	10	10,000
16	**0.1023**	0.0032	0.0124	**0.0000**	0.0010	0.0000	5	0.4	0.2	196	0.001	0.2	10	10,000

Table 9. 2v1 Overall Results

	Number of Blue Wins		
	Offensive	Defensive	Neutral
Benchmark	**16**	5	12
ADP	12	**9**	**14**

Table 9 shows that the ADP generated policies under-perform with respect to the benchmark in the offensive scenario, similar to the 1v1 instance. However, the ADP generated policies do outperform the benchmark in the defensive and neutral scenarios. Sections 4.4.1, 4.4.2, and 4.4.3 respectively explore the results of the ADP generated policies in detail.

4.4.1 Offensive Starts

Similar to 1v1, the section of Blue AUCAVs main goal is to maintain advantage until one can shoot down the Red AUCAV. This task is ideally accomplished in minimal time. We utilize maximizing Blue $\bar{p}_{kill}$ as a proxy for the ability to maintain advantage while minimizing Red $\bar{p}_{kill}$ to avoid positions of disadvantage. Table 10 displays the results from the 20 replications.

Table 10. 2v1 Offensive Performance over 20 Replications

Policy	Blue Won	Time (s)		Blue 1 $\bar{p}_{kill}$		Blue 2 $\bar{p}_{kill}$		Red $\bar{p}_{kill}$	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	16	25.00s	5.99s	0.0011	0.0007	0.0066	0.0022	0.0025	0.0009
ADP, $\pi^{\text{Run 16}}$	12	43.12s	13.21s	0.0012	0.0002	0.0011	0.0006	0.0022	0.0012

Table 10 shows that benchmark policy outperforms the ADP generated policy in the main, secondary, and tertiary statistics, indicating the ADP generated policy is not high-quality. Note the significant difference in time of engagement. The benchmark policy ends an engagement in under 25 seconds on average, whereas the ADP generated policy takes longer than 43 seconds on average with both Blue AUCAVs being shot down more often. The benchmark appears to utilize Blue AUCAV 2 more

often as the attacking aircraft, indicated by the higher $\bar{p}_{kill}$ over replications. The ADP generated policy seems to use a much more balanced approach as both Blue AUCAVs have similar $\bar{p}_{kill}$ values.

4.4.2 Defensive Starts

In a defensive start, the section's first task is to maneuver out of the position of disadvantage. Then, the section can determine maneuvers to gain the position of advantage. In general, the section will want the engagement to last as long as possible when in a position of disadvantage to allow time for the Red AUCAV to make a sub-optimal maneuver or for one of the Blue AUCAVs to have a shot opportunity. Table 11 shows the results from the 20 superlative replications.

Table 11. 2v1 Defensive Performance over 20 Replications

Policy	Blue Won	Time (s)		Blue 1 $\bar{p}_{kill}$		Blue 2 $\bar{p}_{kill}$		Red $\bar{p}_{kill}$	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	5	32.40s	5.90s	0.0012	0.0006	0.0027	0.0019	0.0093	0.0020
ADP, $\pi^{\text{Run 4}}$	**9**	20.32s	10.33s	**0.0042**	**0.0021**	0.0040	0.0024	0.0081	0.0003

Table 11 shows the ADP generated policy outperforms the benchmark policy in two ways. First, the ADP wins more engagements from the defensive start. Second, the ADP generated policy allows the section to gain and maintain a position of advantage in a manner that is better than the benchmark policy as indicated by the Blue 1 $\bar{p}_{kill}$ value. Similar to 1v1 results, the ADP generated policy does not increase the time of the engagement. This is due to the section adopting a high-risk, high-reward strategy where if the section is unable to kill the Red AUCAV on its first pass, it is likely the section will lose the engagement. Figures 29 visualizes a sample engagement in the three-dimensional plane wherein the section is able to shoot down the superior Red AUCAV without losing one of the Blue AUCAVs. Figure 30 shows the angle trade-offs throughout the engagement.

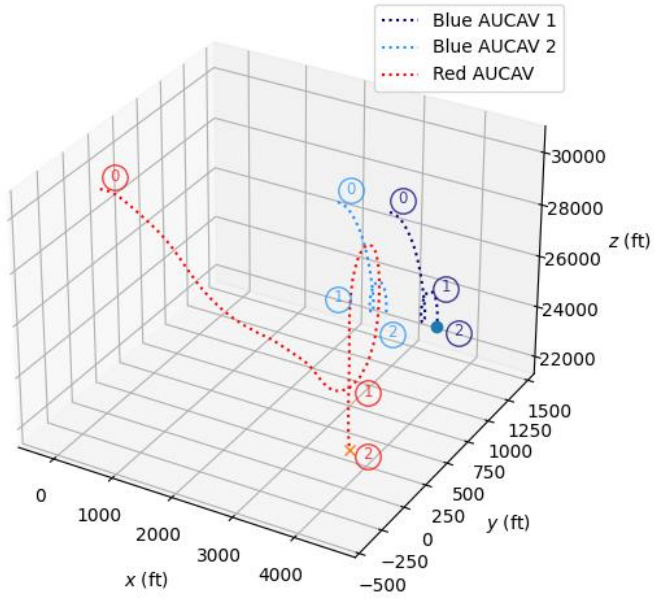

Figure 29. 2v1 **Defensive Engagement in** xyz **Plane**

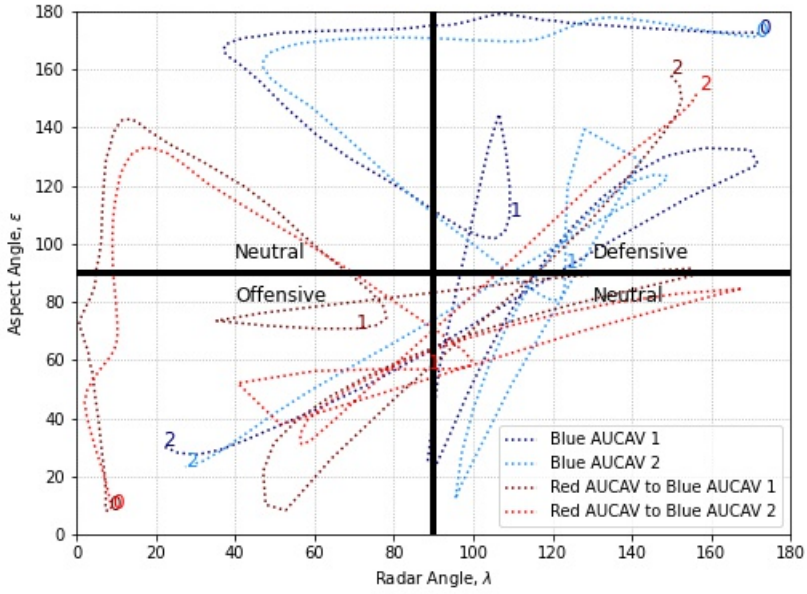

Figure 30. 2v1 **Defensive Engagement: Angle Trade-offs**

Note that Blue AUCAV 1 is identified using a darker, navy blue whereas Blue AUCAV 2 is identified using a lighter, sky blue. There are 3 influential points in this engagement. Influential Point 0 shows the Red AUCAV starting behind the two Blue AUCAVs from a range of about $4,000$ feet, meaning the two Blue AUCAVs are just outside of the Red AUCAV's gun range. As we move from Influential Point 0 to Influential Point 1, the section of Blue AUCAVs performs a tandem maneuver, dropping in altitude while turning to the positive y-axis direction. The Red AUCAV follows while traversing across the xy plane to close in. Figure 30 show this maneuver is able to bring the section out of a position of disadvantage. However, as we approach Influential Point 1, we can see Blue AUCAV 1 is gradually falling back into that position of disadvantage, shown by the darker blue line at Influential Point 1. This is where the ADP generated policy executes the next maneuver.

At Influential Point 1, the section of Blue AUCAVs pulls up initially to travel horizontally across the y-axis direction, and the Red AUCAV responds by pulling up. However, as the Red AUCAV begins to ascend, the section of Blue AUCAVs executes a vertical loop maneuver to slowly loop and face the negative y-axis direction. This is an emergent behavior our section of Blue AUCAVs exhibits, commonly referred to as the split-s (Shaw, 1985). A split-s is used to execute a 180 degree change in direction and is commonly used as disengagement tactic. The success of this disengagement tactic can be seen at Influential Point 1 in Figure 30 where the split-s tactic occurs right after the point and then the navy blue line for Blue AUCAV 1 proceeds from the defensive quadrant to the neutral quadrant below. The Red AUCAV must now loop back under and descend in altitude below the section to travel in the positive y-axis direction towards the section. Finally, at Influential Point 2, the Blue AUCAV has a shot opportunity and is able to kill the Red AUCAV.

4.4.3 Neutral Starts

In a neutral start, the section of Blue AUCAVs must gain advantage over the Red AUCAV quickly. Therefore, the first few maneuvers are significant when it comes to the outcome of the engagement. Time of the engagement is not as important as the ability to gain and maintain advantage over the Red AUCAV. Table 12 shows the results from the 20 superlative replications.

Table 12. 2v1 Neutral Performance over 20 Replications

Policy	Blue Won	Time (s)		Blue 1 $\bar{p}_{kill}$		Blue 2 $\bar{p}_{kill}$		Red $\bar{p}_{kill}$	
		Mean	Half-Width	Mean	Half-Width	Mean	Half-Width	Mean	Half-Width
Benchmark	12	32.40s	5.90s	0.0051	0.0020	0.0049	0.0019	0.0040	0.0013
ADP, $\pi^{\text{Run 7}}$	14	26.21s	3.14s	0.0061	0.0047	0.0051	0.0036	0.0026	0.0012

Table 12 shows the ADP generated policy outperforms the benchmark policy in terms of sheer number of wins for the section of Blue AUCAVs. The confidence intervals for the Blue AUCAV 1 and Blue AUCAV 2 $\bar{p}_{kill}$ quantities are much wider when using the ADP generated policy than with the benchmark. The reason is the ADP generated policy takes on a high-risk, high-reward strategy where if the Red AUCAV is not killed on the first pass of shot opportunities, the likelihood the section wins the engagement drops dramatically. Similar to 1v1 results, the ADP generated policy does not significantly outperform the benchmark in the neutral start. There are no noticeable changes to the time an engagement will last, the ability to maintain advantage, or the ability to avoid positions of disadvantage. Figures 31 and 32 show a sample engagement in the xyz space and xz plane from a neutral start. Figure 33 shows the angle trade-offs throughout the engagement.

Influential Point 0 shows the starting state of the engagement where the section is facing in the negative x-axis direction while the Red AUCAV is facing in the positive x-axis direction. All AUCAVs in the engagement begin a gradual descent in altitude. At Influential Point 1, Blue AUCAV 1 initiates a slightly tighter turn than

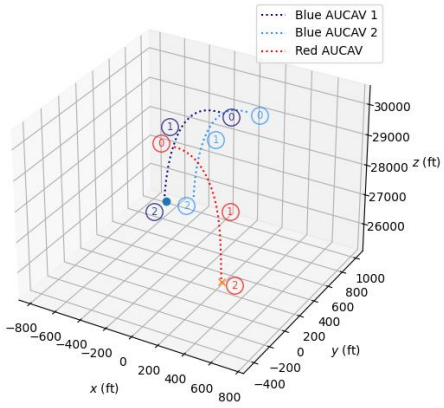

Figure 31. 2v1 **Neutral Engagement in** xyz **Plane**

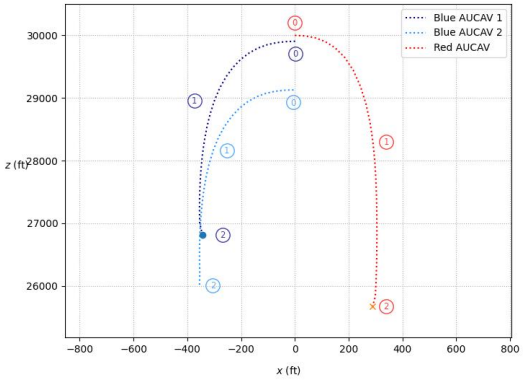

Figure 32. 2v1 **Neutral Engagement in** xz **Plane**

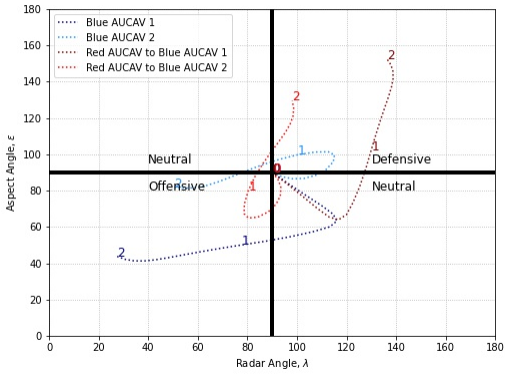

Figure 33. 1v1 **Neutral Engagement: Angle Trade-offs**

its counterpart, Blue AUCAV 2. As this is happening, the Red AUCAV is focusing on Blue AUCAV 2 according to its position-only policy. At Influential Point 3, the Red AUCAV is starting to turn in the negative x-axis direction to aim at Blue AUCAV 2. As this is happening, the Blue AUCAV 1 is setting up position to have the Red AUCAV within its 30° radar angle WEZ due to its tighter turn at Influential Point 1. The turn continues until Blue AUCAV 1 earns a shot opportunity and is successful. Figure 33 shows the ability of the ADP generated policy to not allow the Red AUCAV to gain a position of advantage over any aircraft in the section.

While this tactic is not an exact replica of common fighter tactics, the generic strategy is reminiscent of a defensive maneuver called the half-split. The half-split typically occurs in a horizontal plane where a section is flying abreast and is visualized in Figure 34. As an attacker comes into frame, the inside fighter executes a turn while the outside fighter continues straight before turning. If the attacker pursues the inside fighter, the outside fighter initiates a slightly tighter turn to earn a shot opportunity as the attacker focuses on the inside fighter.

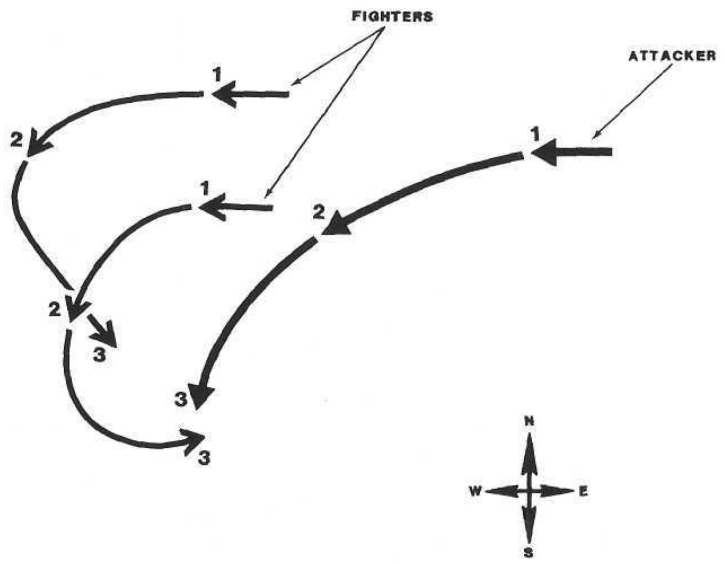

Figure 34. The Half-Split (Shaw, 1985)

The ADP generated policy executes a variant of this tactic except in the vertical

dimension and neutral start. The section initiates a gradual descent down while the Red AUCAV follows and starts to pursue Blue AUCAV 2, which is the inside fighter in this scenario. As the Red AUCAV focuses on Blue AUCAV 2, Blue AUCAV 1 is turning slightly tighter and able to earn that shot opportunity at Influential Point 3. This is not an exact replica of the half-split and should not be considered emergent behavior. However, it is salient to note the similar tactics between the ADP generated policies and common fighter tactics, despite being in different dimensions.

V. Conclusions and Recommendations

This research examines the air combat maneuvering problem (ACMP) in one-versus-one (1v1) and two-versus-one (2v1) engagements in the pursuit of generating individual and section policies respectively that are high quality. We formulate a discrete time, infinite horizon Markov decision process (MDP) model, which is the first to generically model the ACMP in terms of M Blue AUCAVs and N Red AUCAVs, called the MvN ACMP. We also formulate the first directed energy weapon (DEW) within the ACMP but were unable to test due to resource constraints. Finally, with regards to the Red AUCAV's decision making ability, we do not use the canonical mini-max algorithm for the Red AUCAV but instead leverage the high-quality benchmark policies found in recent years (Crumpacker, 2021; Wang et al., 2016; Fang et al., 2016; McGrew et al., 2010).

The continuous nature of our state space combined with the high-dimensionality does not allow for an optimal policy to be computed using exact dynamic programming. An approximate dynamic programming (ADP) solution procedure is formulated to determine a high-quality policy using an approximate policy iteration strategy where the value function is approximated using a neural network. We test for multiple activation functions, layers, and activation units within each layer to determine the highest quality architecture for our problem instances. We expand upon the novel state sampling scheme presented by Crumpacker (2021) to sample from in front of and behind the Red AUCAV, which helps dissipate the large positive and negative contributions throughout the state space and combat the sparse rewards problem.

We analyze the MvN ACMP and ADP solution procedure in 1v1 and 2v1 cases across three problem instances defined by the starting states of the engagement: offensive, defensive, and neutral. To evaluate performance of the resulting policies, we utilize the position-energy benchmark policy to determine quality. We compare per-

formance of the ADP generated policies to the position-energy policy across offensive, defensive, and neutral starts in the 1v1 and 2v1 case, a total of 6 problem instances.

The results of our computational experiments show we truly outperform the benchmark policy on 4 out of the 6 problem instances. These results show the need for future research to consider the offensive, defensive, and neutral starts as their own problem. The idea that one single policy can be used at all times across an air combat engagement is not feasible. Pope et al. (2021), in what is arguably the highest quality research on this topic to date, determine a solution approach that defeated a USAF test pilot five to zero in simulated engagements. One of the main features of their research was to break down the problem of air combat into a hierarchy of tasks and sub-tasks, an approach called hierarchical reinforcement learning. We believe the results of our computational experiments directly support Pope et al. (2021) and the need for hierarchical approach in air combat.

Qualitatively, the ADP generated policies elicit some common aerobatic maneuvers, such as the split-s and the half-split. Despite this emergence, the resulting policies did not exactly mimic human fighter pilot behavior as many of the engagements explored changes in altitude more than changes in lateral position, which is typically opposite of human fighter pilot behavior. In fact, the ADP generated policies mimicked the half-split tactic in the xz plane instead of the xy plane. This vertical behavior may be a more effective tactic when there is no human pilot.

In the 1v1 ACMP, future research can build on the DEW model presented in the MvN ACMP and explore what tactics fighter pilots should fly given a DEW on the weapon system. Moreover, future research can focus extensively on the three starting positions and determine a much higher quality policy for that starting position. Then, once these three starting position policies are determined, future research can then determine a higher-level policy for when to implement each high-quality policy. It is

clear the future of this research lies in hierarchical approach. As an initial step, the angle-tradeoff graphs presented in this research may present a benchmark policy to label flight status.

In the 2v1 ACMP, future research can explore various communication schemes. This research assumes a centralized communication scheme where the section senses and acts as one. However, this has scalability issues as the cardinality of our action space jumps to over $324,744$ in the 2v1 case only. The issues worsens when we expand into 4v2 and higher. Future research can explore decentralized communication using partially observable MDPs (PO-MDPs) where each AUCAV only sees part of the state space. The idea of controlling multiple entities also lends itself to transfer learning, which seeks to determine how to learn quicker and smarter from pre-existing agents and policies. All of these topics are of value and will have near-term applications for the United States Air Force.

Bibliography

Barr, R. S., Golden, B. L., Kelly, J. P., Resende, M. G. and Stewart, W. R. (1995), 'Designing and reporting on computational experiments with heuristic methods', *Journal of Heuristics* **1**(1), 9–32.

Burgin, G. H. and Owens, A. (1975), 'An adaptive maneuvering logic computer program for the simulation of one-to-one air-to-air combat. volume 2: Program description'.

Burgin, G. H. and Sidor, L. (1988), Rule-based air combat simulation, Technical report, Titan Systems Inc La Jolla CA.

Canese, L., Cardarilli, G. C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M. and Spanò, S. (2021), 'Multi-agent reinforcement learning: A review of challenges and applications', *Applied Sciences* **11**(11), 4948.

Choudhury, M., Aoun, A. and Badaway, D. (2019), 'Final Report of the Panel of Experts on Libya established pursuant to Security Council Resolution 1973 (2011)', *United Nations* .

Crumpacker, J. B. (2021), Air combat maneuvering via operations research and artificial intelligence methods, Master's thesis, Air Force Institute of Technology.

Fang, J., Zhang, L., Fang, W. and Xu, T. (2016), Approximate dynamic programming for cgf air combat maneuvering decision, *in* '2016 2nd IEEE International Conference on Computer and Communications (ICCC)', IEEE, pp. 1386–1390.

Fayjie, A. R., Hossain, S., Oualid, D. and Lee, D.-J. (2018), Driverless car: Autonomous driving using deep reinforcement learning in urban environment, *in* '2018 15th International Conference on Ubiquitous Robots (UR)', pp. 896–901.

Glorot, X. and Bengio, Y. (2010), Understanding the difficulty of training deep feedforward neural networks, *in* 'Proceedings of the thirteenth international conference on artificial intelligence and statistics', JMLR Workshop and Conference Proceedings, pp. 249–256.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P. et al. (2018), 'Soft actor-critic algorithms and applications', *arXiv preprint arXiv:1812.05905* .

Jenkins, P. R., Robbins, M. J. and Lunday, B. J. (2021), 'Approximate dynamic programming for military medical evacuation dispatching policies', *INFORMS Journal on Computing* **33**(1), 2–26.

Ma, X., Xia, L. and Zhao, Q. (2018), Air-combat strategy using deep q-learning, *in* '2018 Chinese Automation Congress (CAC)', IEEE, pp. 3952–3957.

Makar, R., Mahadevan, S. and Ghavamzadeh, M. (2001), Hierarchical multi-agent reinforcement learning, *in* 'Proceedings of the fifth international conference on Autonomous agents', pp. 246–253.

Mazarr, M. J., Blake, J. S., Casey, A., McDonald, T., Pezard, S. and Spirtas, M. (2018), *Understanding the Emerging Era of International Competition: Theoretical and Historical Perspectives*, RAND Corporation, Santa Monica, CA.

McGrew, J. S., How, J. P., Williams, B. and Roy, N. (2010), 'Air-combat strategy using approximate dynamic programming', *Journal of Guidance, Control, and Dynamics* **33**(5), 1641–1654.

Mishkin, D. and Matas, J. (2015), 'All you need is a good init', *arXiv preprint arXiv:1511.06422* .

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K. and Grossman, D. (2020), *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*, RAND Corporation, Santa Monica, CA.

Pope, A. P., Ide, J. S., Micovic, D., Diaz, H., Rosenbluth, D., Ritholtz, L., Twedt, J. C., Walker, T. T., Alcedo, K. and Javorsek, D. (2021), 'Hierarchical reinforcement learning for air-to-air combat', *arXiv preprint arXiv:2105.00990* .

Powell, W. B. (2011), *Approximate Dynamic Programming: Solving the curses of dimensionality*, Vol. 703, 2nd edn, John Wiley & Sons.

Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons.

Shaw, R. L. (1985), *Fighter Combat*, United States Naval Institute, 291 Wood Road, Annapolis, MD, 21402.

Sutton, R. S. and Barto, A. G. (2018), *Reinforcement Learning: An Introduction*, 2 edn, MIT press, Cambridge, MA.

Tan, M. (1993), Multi-agent reinforcement learning: Independent vs. cooperative agents, *in* 'Proceedings of the tenth international conference on machine learning', pp. 330–337.

Toubman, A. (2020), Calculated Moves: Generating Air Combat Behaviour, PhD thesis, Leiden University.

United States Department of Defense (2018), *National Defense Strategy*, Washington, DC.

United States Department of the Air Force (2021), *Air Force Doctrine Publication 1*, Washington, DC.

Virtanen, K., Karelahti, J. and Raivio, T. (2006), 'Modeling air combat by a moving horizon influence diagram game', *Journal of Guidance, Control, and Dynamics* **29**(5), 1080–1091.

Wang, F.-Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., Zhang, J. and Yang, L. (2016), 'Where does alphago go: From church-turing thesis to alphago thesis and beyond', *IEEE/CAA Journal of Automatica Sinica* **3**(2), 113–120.

Wang, M., Wang, L., Yue, T. and Liu, H. (2020), 'Influence of unmanned combat aerial vehicle agility on short-range aerial combat effectiveness', *Aerospace Science and Technology* **96**, 105534.

Wiliam, D. (2006), 'The half-second delay: what follows?', *Pedagogy, Culture & Society* **14**(01), 71–81.

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y. and Jiang, T. (2019), 'Deep reinforcement learning for smart home energy management', *IEEE Internet of Things Journal* **7**(4), 2751–2762.

Zohuri, B. (2016), *Directed Energy Weapons*, Springer.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
24–03–2022	Master's Thesis	August 2020 – March 2022

4. TITLE AND SUBTITLE

Team Air Combat Using Model-based Reinforcement Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Mottice, David A., 2 Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-22-M-157

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Strategic Development Planning & Experimentation Office
Mr. David M. Panson
1864 4th Street
Wright-Patterson AFB, OH 45433
(937) 904-6539

10. SPONSOR/MONITOR'S ACRONYM(S)

SDPE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release, distribution is unlimited.

13. SUPPLEMENTARY NOTES

This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

We formulate the first generalized air combat maneuvering problem (ACMP), called the MvN ACMP, wherein M friendly AUCAVs engage against N enemy AUCAVs, developing a Markov decision process (MDP) model to control the team of M Blue AUCAVs. The MDP model leverages a 5-degree-of-freedom aircraft state transition model and formulates a directed energy weapon capability. Instead, a model-based reinforcement learning approach is adopted wherein an approximate policy iteration algorithmic strategy is implemented to attain high-quality approximate policies relative to a high performing benchmark policy. The ADP algorithm utilizes a multi-layer neural network for the value function approximation regression mechanism. One-versus-one and two-versus-one scenarios are constructed to test whether an AUCAV can outmaneuver and destroy a superior enemy AUCAV. The performance is evaluated across offensive, defensive, and neutral starts, leading to 6 problem instances. The ADP policies outperform the position-energy benchmark policy in 4 of 6 problem instances. Results show the ADP approach mimics certain basic fighter maneuvers and section tactics.

15. SUBJECT TERMS

markov decision processes, reinforcement learning, artificial intelligence, air combat

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Matthew J. Robbins, AFIT/ENS
U	U	U	UU	104	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4539; Matthew.Robbins@afit.edu