Optimizing Cooperative Games for Cognitive Communication UAVs with Q-Learning

Mark Rahmes, David Chester, Richard Clouse, Jodie Hunt
Harris Corporation, Space and Intelligence Systems
Melbourne, Florida 32904
mrahmes@harris.com, dchest04@harris.com, rclouse@harris.com, jhunt11@harris.com

ABSTRACT

Currently, distributed communications networks based on multiple unmanned aerial vehicles (UAVs) are limited in terms of reliability and network availability. The capacity for each UAV to serve as a node in the network is constrained by limited energy stores, dynamic changes in the network topology, and latency/jitter issues. Typical approaches to address these challenges have focused on partitioning of the network to work around the failed nodes, but the attendant degraded communications links and lengthy network outages underscore the need for a better solution. An innovative approach based on the use of a self-forming, self-organizing, cooperative, autonomous system of distributed UAV communication nodes is being investigated. By enabling each UAV to act collectively and cooperatively, a multi-UAV network's communication links can be made more resilient, resulting in enhanced levels of network availability and improved service quality. To achieve this, we investigated the concept of opportunistic arrays to aid in the development of a cooperative, cognitive system encompassing multiple vehicles. Based on simulations, we have also been able to demonstrate that optimal vehicle positions can be directed using decision algorithms that embody elements of game theory. In addition, by implementing a cooperative reasoning engine for system-level oversight or harmonization, we were able to ensure optimal performance of the overall system and achieve enhanced levels of service quality based on multiple measures of effectiveness.

ABOUT THE AUTHORS

Dr. Mark Rahmes has 25 years of experience at Harris Corporation as an Electrical/Computer Engineer and Senior Research Scientist. He earned his BSEE from The Citadel, MSEE from Duke University and PhD in Operations Research from Florida Tech. Dr. Rahmes is a retired U.S. Navy Reserve Captain and served 22 years as a Commanding Officer, Engineering Duty Officer and Surface Warfare Officer. He currently has 48 patents granted and has published 51 professional publications. At Harris, Dr. Rahmes serves as a Principal Investigator and Chief Engineer.

Dr. David Chester has 34 years of experience, 29 at Harris, serving in various roles, including Systems/Project Engineer, Lead Applications Engineer, and Chief Architect. Primary technology focus areas include Cognitive Network Technologies, Communications and Digital Signal Processing, and Communications Theory and Systems. His education encompasses PhD ECE and MSEE, University of Cincinnati; BS Physics, Xavier University; Hauck Research Fellow and he has over 80 professional publications and 53 U.S. patents granted.

Richard Clouse has over 34 years of experience, 2 years at Harris, serving in various technical and managerial roles. He is currently Director of Strategy and Technology where his primary responsibility is planning and management of a broad portfolio of technologies. His current focus is application of Deep Learning to EW systems. He earned his BSEE from Northeastern University, MSEE from Polytechnic Institute of New York and MBA from New York University. He has written and presented multiple papers, is an IEEE region 1 award recipient and patent holder.

Jodie Hunt has 10 years of technical experience in analytical chemistry and optoelectronics. She earned a BS and MS in Chemistry from Wright State University in Dayton, Ohio. At the Air Force Research Laboratories at Wright Patterson-AFB, she spent five years at the Materials Directorate working on epitaxial growth of III-V semiconductor materials and five years at the Sensors Directorate working on the characterization and fabrication of numerous device materials for sensor applications. She has numerous publications and conference papers.

Optimizing Cooperative Games for Cognitive Communication UAVs with Q-Learning

Mark Rahmes, David Chester, Richard Clouse, Jodie Hunt
Harris Corporation, Space and Intelligence Systems
Melbourne, Florida 32904
mrahmes@harris.com, dchest04@harris.com, rclouse@harris.com, jhunt11@harris.com

INTRODUCTION

UAVs may be used for communication relaying. Information must often be transmitted to a base station in real time. However, limited communication ranges and free line of sight may make direct transmissions from distant targets impossible. This problem can be solved using relay chains consisting of one or more intermediate relay UAVs. This leads to the problem of positioning such relays given known obstacles which may be located using a digital terrain model, while taking into account a possibly mission-specific quality measure, which may be determined by free space path loss formula. The maximum quality of a chain may depend strongly on the number of UAVs allocated (Burdakov, O., 2010).

One important reason to model and simulate where to position UAVs is to help determine the number required and how much power each should be able transmit and receive across a desired distance. It is important for communication nodes in a chain to transmit information reliably from distant targets back to a base station for military purposes. If there are any degradations within the system, we want to be able to determine the operating boundaries.

A relay network is a broad class of network topology commonly used in wireless networks and is characterized by a source and destination connected by intermediate nodes. This addresses the fact that, in a widely distributed network, the source and destination cannot communicate directly because the link would be degraded due to, for example, distance-dependent free space path loss (FSPL) or signal interference. The intermediate nodes in a relay network allow for shorter signal propagation distances that effectively mitigate the impact of FSPL and RF interference. In information theory, a relay channel is a probability model of the communication parameters between a sender and a receiver aided by one or more intermediate relay nodes. Using cognitive processing to assess these parameters, we can achieve optimal positioning of an autonomous UAV swarm. The UAV-positioning algorithm uses various input parameters, such as distance and channel quality, to guide the control decisions. A digital terrain model (DTM) can also be used as input, as shown in Figure 1, to ensure unobstructed line-of-sight signal paths.

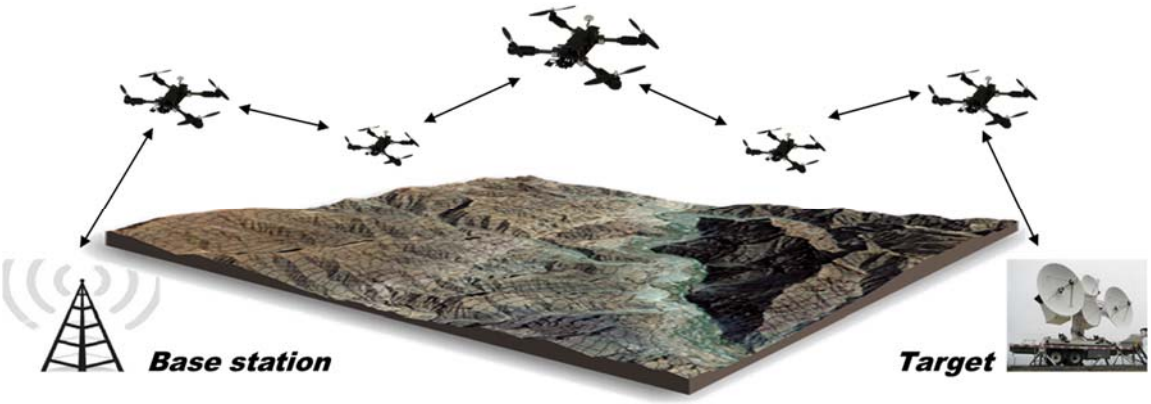

Figure 1. Notional Multi-UAV Communication Relay Network, with Digital Terrain Model (DTM) used as an input Parameter

Improvements in overall system performance and an enhanced capability to meet mission requirements are just some of the benefits that can be realized through the cooperative self-positioning of UAVs. An inherent feature of the cognitive processing required for autonomous cooperative UAV positioning is detailed insight into the hierarchical

elements of the system, i.e., the physical layer, application layer, packet layer, etc. That insight can be exploited at the system level to realize additional improvements in system performance. In the major sections of the paper that follow, we present a methodology for achieving these system performance improvements as part of our discussions on the system game model, details of the simulation scenarios, a method for optimizing each system parameter through a training phase, and the projected performance benefits.

SYSTEM GAME MODEL

Game theory is the study of strategic decision-making and mathematical modeling of conflict and cooperation between intelligent, rational decision-makers, and is often thought of as an interactive decision theory. It has been applied to economics, political science, psychology, logic, biology and other complex issues. Modern game theory began with the idea of the existence of mixed-strategy equilibrium in two-person zero-sum games, applied to economics. Later, this evolved to provide a theory of expected utility, which allowed mathematicians and economists to treat decision-making with uncertainty. The notion of probabilistic predictions utilizing game theory is critical to many decision-making applications because optimizing user experience requires being able to compute expected utilities of mutually exclusive data. We are using game theory as an optimal decision algorithm to choose best direction for UAVs to travel to their assigned positions within a communication relay.

The multi-UAV system is dynamic and flexible, enabling it to respond to emerging conditions that could impact network performance. Its ability to dynamically adjust the individual parameter weightings to account for changing conditions ensures that UAV positioning decisions take into account both the immediate needs of a given network node and the overall needs of the network. These conflicting needs form the basis for a game scenario in which each UAV is alternately pushed or pulled over time to new positions until reaching a steady state, or Nash Equilibrium.

A high-level multi-UAV system block diagram is presented in Figure 2. As indicated in the figure, individual parameter weights are initially optimized through training. Once the parameter weights are determined, the system uses them to support control decisions that optimize UAV positioning.

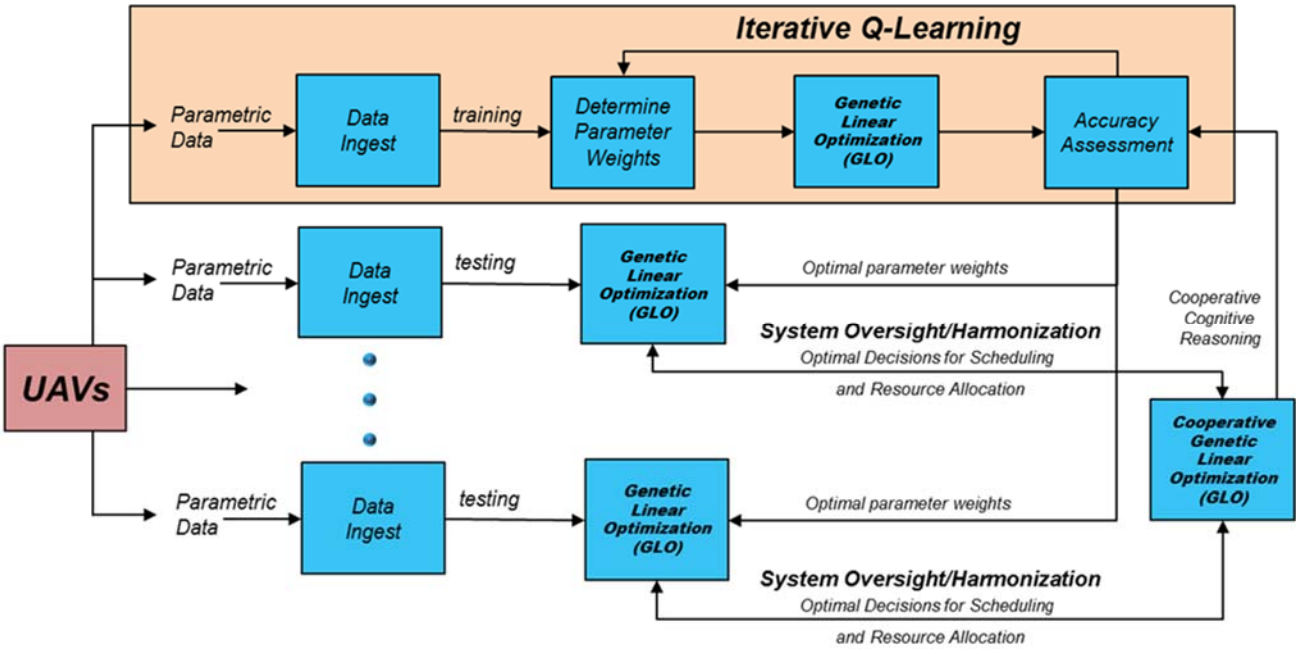

Figure 2. Functional Block Diagram

One good implementation of game theory is to use Linear programming to define a UAV positioning decision matrix. Each row of the matrix corresponds to one of 27 directions in which to move, e.g., north, east, south, west, up, down, as shown in Figures 3 and 4. Each column is associated with a parameter, e.g., transmit power, reception quality for nearby UAVs, distance to nearby UAVs and range to initial optimal assigned position. The last parameter serves to keep the UAV near this position.

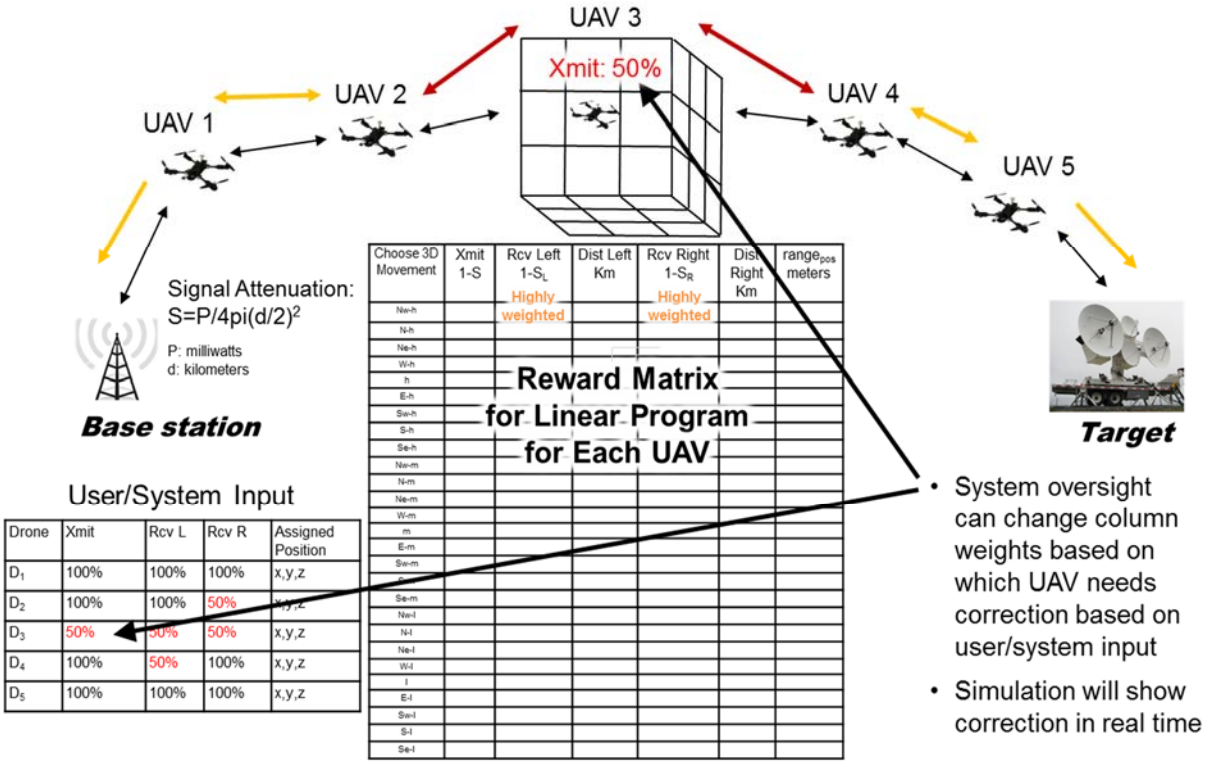

Figure 3. Scenario 1

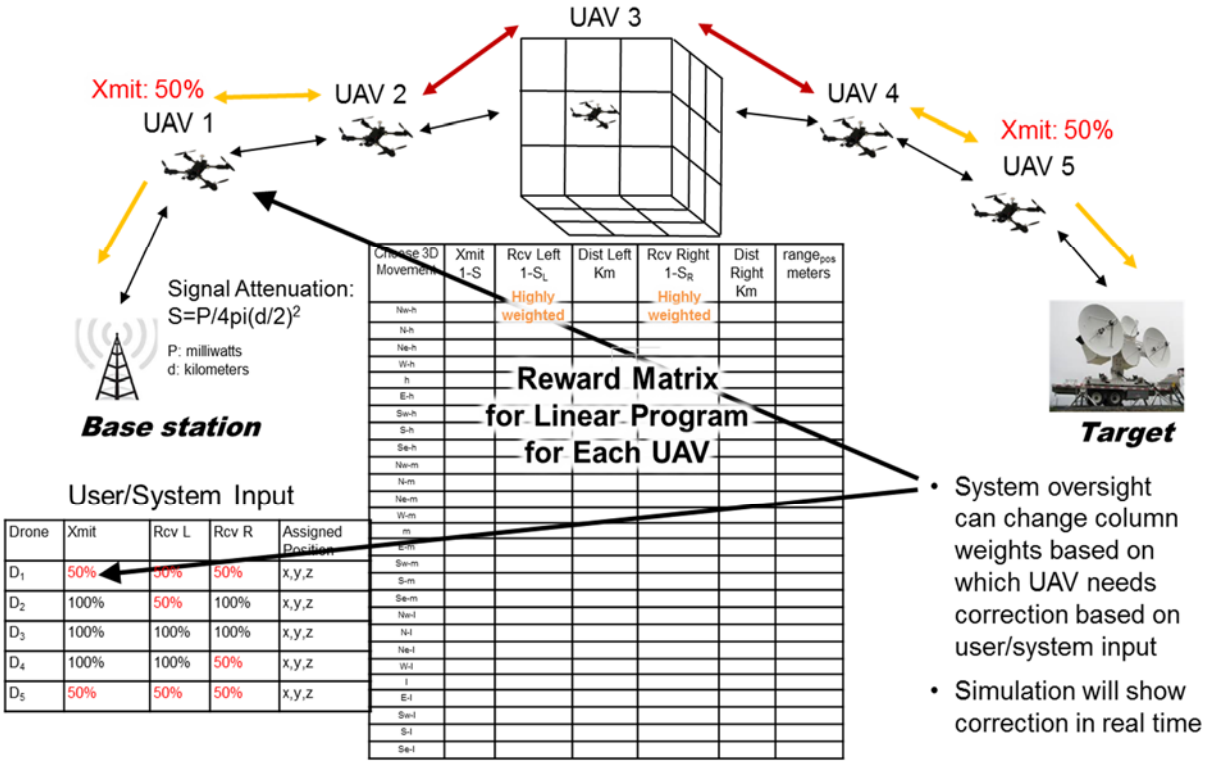

Figure 4. Scenario 2

The use of genetic algorithms (GA) is currently a very popular approach to optimally determining the best characteristic values for a given situation. GA can play an important part in creating test scenarios to determine optimal values for multiple degrees of freedom. Analysts can also use game theory to develop models that better predict actual behavior. The innovative approach being presented here is to combine GA and linear programming to create an enhanced game theory-based decision algorithm, which we refer to as Genetic Linear Optimization (GLO). We handle the genetics portion of the algorithm with an iterative process.

In this paper, we focus on two experiments that we conducted to demonstrate the benefits of cooperation. The weighting of parameters is the essential driver for positioning decisions that move each UAV such that positive overall system performance is achieved. For the results described here, five UAVs were used as communication relays. The system keeps a status of the health of each UAV and adjusts the parameter weights, as needed, to ensure that ongoing UAV positioning instructions continue to maintain optimal system communication links.

Linear optimization is useful for solving game theory problems and finding optimal strategies. In a classic game theory scenario, a conductor must choose among a set of actions, the consequences of which depend on either certain states about which the conductor is not completely informed (i.e., subjective uncertainty) or the result of random, independent processes (i.e., objective uncertainty) (Roy, O., 2010). The "expected utility" forms the basis for a prediction of what the conductor will choose in this uncertain environment. The use of probabilistic predictions and game theory is an essential element of many decision-making applications, given the need to compute expected utilities for mutually exclusive objectives to optimize performance. The Nash Equilibrium is synonymous with objective function, or value, of the game in an integer linear program.

We have chosen to implement a game theoretic solution using linear programming optimization. Our implementation handles non-linear problems via the use of subset summing of all feasible solutions during the decision-making portion of the algorithm. Other possible solutions include deep learning or neural networks, but require a large amount of training data. We use an interior-point algorithm, i.e., the primal-dual method, which must be feasible for convergence. The primal standard form, which is used to calculate optimal tasks and characteristics (Winston, W., 2003), is:

$$minimize\ (f * x)\ s.t. \quad A * x = b, \quad x \geq 0 \tag{1}$$

The dual problem, which is used to calculate optimal parameters, is:

$$maximize\ (b' * y)\ s.t. \quad A' * y + s = f, \quad s \geq 0 \tag{2}$$

Since we know the optimal direction decision based on detected UAV health, we can find the associated parameter (column), given the decision (row). We then use an error function to determine the optimal weights and the importance of each parameter for Q-Learning.

We have chosen to determine the optimal weights for each input parameter using Q-Learning mathematical method which requires a relatively small amount of training data to be successful. An additional benefit of Q-Learning is that it can also be used to model the response time in the dynamic system when the system oversight portion is used to direct individual UAVs to a new desired optimal position due to changing environmental system parameters.

SIMULATIONS

Free-space path loss (FSPL) is the loss in signal strength of an electromagnetic wave with an unobstructed line-of-sight path through free space (usually air). We model signal attenuation using the inverse square law:

$$S = P/(4 \pi d^2) \tag{3}$$

where S is the power spatial density in watts per square meter; P is the equivalent isotropically radiated power in watts referenced to 1 meter.

The object of the game is to minimize the path loss for each link in the communication relay chain such that an overall higher quality of communication is achieved, ideally a level of quality that ensures that all mission requirements will be met. The dynamics of two scenarios are discussed later in this paper. The first scenario involves a degradation of UAV 3. The second involves a degradation of UAVs 1 and 5. Note that degradation in this context is understood to mean reduced throughput as a function of propagation distance and not a full node failure. The simulation demonstrates that UAVs can reposition themselves closer to degraded UAV(s) to regain the minimum required network throughput.

The oversight block modifies the parameter weights to optimize UAV positioning. It should be noted that there can be multiple levels of hierarchy for system oversight or harmonization. Also, training of the oversight block with regard to parameter weights may need to be mission-specific if the concept of operations (CONOPS) changes significantly.

One mission specific application for our system is data exfiltration. The use of game theory as a computation engine scales well with multiple UAVs. The number of parameters per UAV remains constant since only the closest neighbors are considered. A simulation tool allows the user to visualize and record the time for each UAV to arrive at optimal station which increases when more UAVs are added to the chain. Sensor data can be exfiltrated by one or more UAVs that act as data mules by visiting each sensor in order to establish a communication link (Klein, D., 2010).

Q-LEARNING

We use Q-Learning as a method to figure out the best parameter weights for optimal system performance. Q-Learning attempts to learn the relevance of each parameter, previously discussed, such that the best decision is chosen. This method requires sufficient training data in order for the system to learn under various conditions. Once the parameter weights are learned, the system will then perform best.

We additionally allow the oversite module to adjust parameter weights for individual UAVs in order to enhance end to end system performance. The use of simulations allow for a valuable tool for a user to visually check how the system responds to environmental perturbations. The learning rate can be further adjusted by the user to understand functional time lines in the system.

Q-Learning can be used for function approximation, which enables the application of the algorithm to large problems, even when state space is continuous and, therefore, infinitely large (van Hasselt, H., 2012). The solution described here is characterized by the system learning what the weights should be for each parameter to achieve optimal system performance.

The system uses Q-Learning to assign optimal weights as a Markov decision process (MDP). Q-Learning is useful in determining weights for each dimension in the system and can provide insight into relationships among objectives, which improves the understanding of the problem. Although each period is modeled as independent, four dimensions within a period are considered dependent, sequential Markov functions (Parisi, S., 2014).

We use an optimal, multi-objective Markov action-selection decision-making function with Q-Learning. The system is flexible and can handle any number of inputs. By optimizing weights from these inputs, multiplied by the Nash Equilibrium (NE) values for each of the dimension possibilities per period, optimization of UAV positioning is achieved, which results in significant improvements in overall system performance. Our model features six dimensions with conflicting objectives that depend on each UAV to achieve an optimal positioning solution.

When using a mathematical model to characterize dynamic real-world processes, it becomes necessary to use approximations to reduce the otherwise intractable complexities. Linearity assumptions are usually sufficient approximations. Other important approximations are needed to account for uncertainties associated with the accuracy and/or completeness of the data input to the model (Demers, A., Keshav, S., Shenker, S., 1989).

Without sufficient insight into the relevant information, it becomes necessary to assign approximate values to variables in a linear equation. Moreover, that information may change. Sensitivity analysis, i.e., a systematic assessment of the

extent to which solutions are sensitive to changes in data (Demers, A., Keshav, S., Shenker, S., 1989), is essential in quantifying the validity of the results. In our example, we define the Q-Learning equation as:

$$\text{newWeight} = (1\text{-errorA,B,C,D,E\_norm}) * (\text{NashEquilibriumValue}) / (\sum\_\text{vparams} [(1\text{-errorA,B,C,D,E\_norm}) * (\text{NashEquilibriumValue})]) \quad (4)$$

where A, B, C, D and E are parameters.

$$\text{QLearnedWeight} = \text{oldWeight} + \text{learningRate} * (\text{newWeight} - \text{oldWeight}) \quad (5)$$

The results of the simulations for the two scenarios described above are shown in Figures 5, 6, 7, and 8 for two different learning rates for each scenario. The parameter weights are normalized such that they add up to one. From this it can be determined which parameters are most important for optimal settings. Tuning the parameter weights is also useful in providing valuable insight into the system (Cohen, R., Rahmes, M., Fox, K., Lemieux, G., 2016).

The learning rate determines to what extent newly acquired information will override old information. A factor of 0 will disable learning by an agent, while a factor of 1 will force the agent to consider only the most recent information. The discount factor determines the importance of future rewards. A factor of 0 will make an agent short-sighted by only considering current rewards, while a factor approaching 1 will make an agent strive for a long-term higher reward.

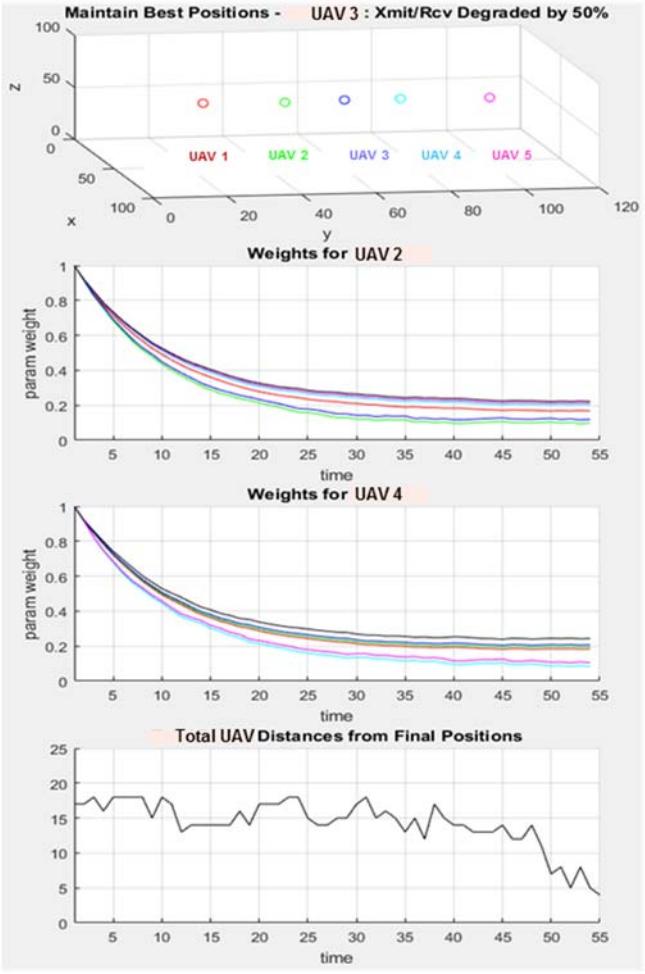

Figure 5. Scenario 1, Learning Rate = 0.1

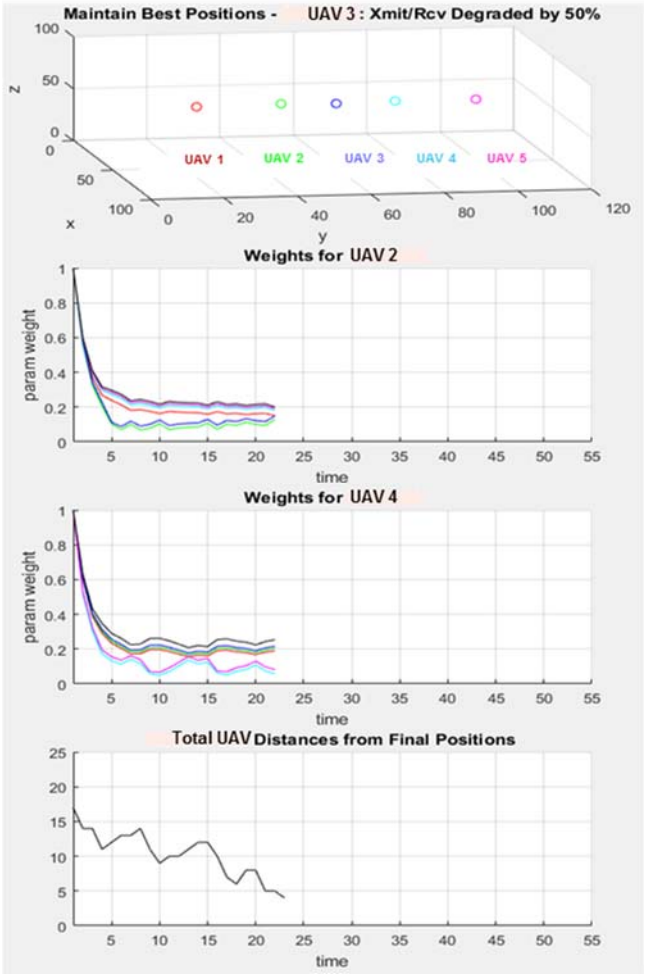

Figure 6. Scenario 1, Learning Rate = 0.5

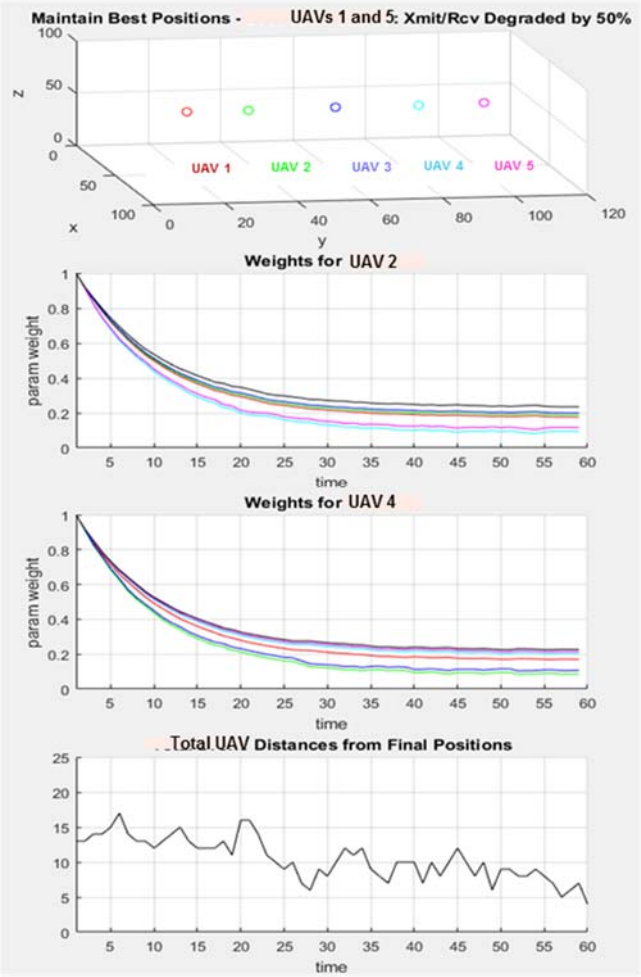

Figure 7. Scenario 2, Learning Rate = 0.1

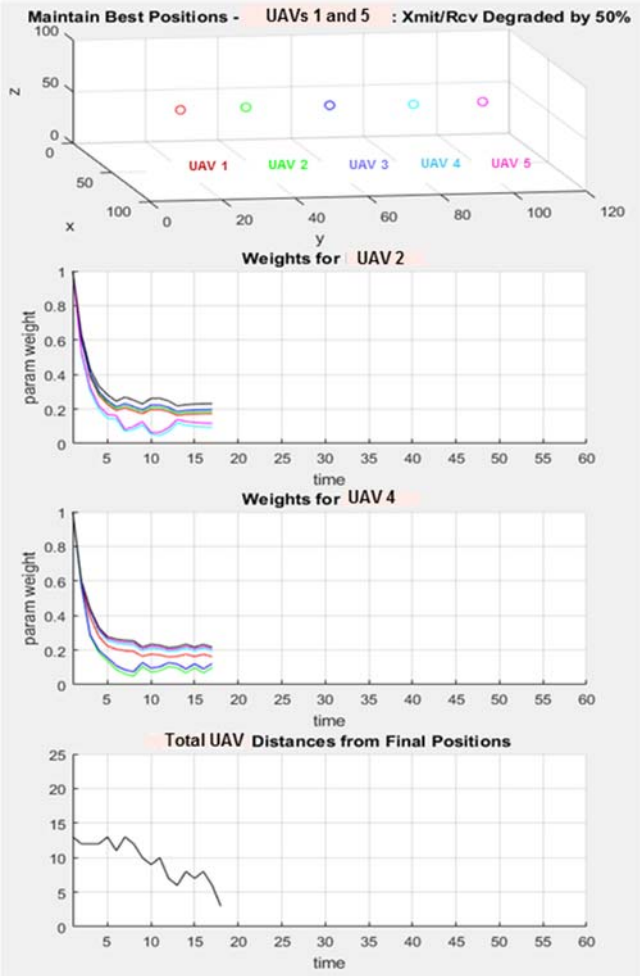

Figure 8. Scenario 2, Learning Rate = 0.5

Note that in Scenario 1, UAVs 2 and 4 move closer to UAV 3 to provide assistance. In Scenario 2, UAVs 2 and 4 move closer to 1 and 5 respectively to provide a higher throughput relay link, which is the goal. For each UAV, the six weights are optimally determined using Q-Learning and each has a unique color assigned in the figures.

Note that the dynamic maneuver is completed in 55 turns when the learning rate is 0.1 and in 23 turns when the learning rate is 0.5 for Scenario 1.

GLO Oversight causes Rcv Left and Dist Left parameters to receive lower weighting for UAV 2, while Rcv Right and Dist Right receive lower weighting for UAV 4.

Note that the dynamic maneuver is completed in 60 turns when the learning rate is 0.1 and in 18 turns when the learning rate is 0.5 for Scenario 2.

GLO Oversight causes Rcv Right and Dist Right parameters to receive lower weighting for UAV 2, while Rcv Left and Dist Left receive lower weighting for UAV 4.

RESULTS

In our experiments, we assumed an Additive White Gaussian Noise (AWGN) channel. We used Quadrature Phase Shift Keying (QPSK) and assumed a necessary raw (before code correction) bit error rate (BER) of at least 10^{-4}. Furthermore, based on theory, we assumed an operational throughput of 0.5 dB. We chose a P of 20 watts for equation (3) and an in-band noise floor of 2 watts at each UAV. This yielded a signal-to-noise ratio (SNR) as a function of distance. Degradation could be due to a local increase in the noise floor.

The UAV swarm needs to arrange itself such that the SNR is maintained at or above the level where the Eb/N0 yields the required BER on each link. Once the required throughput is met, the minimum transmit power can be optimized.

Figure 9 shows that both scenarios 1 and 2 can handle the degradation of one or two UAVs when using cooperation. However, the experiments also indicate that it is unlikely that the system could handle additional UAV degradation(s). Without cooperation, neither scenario can achieve the required signal integrity for all communication links.

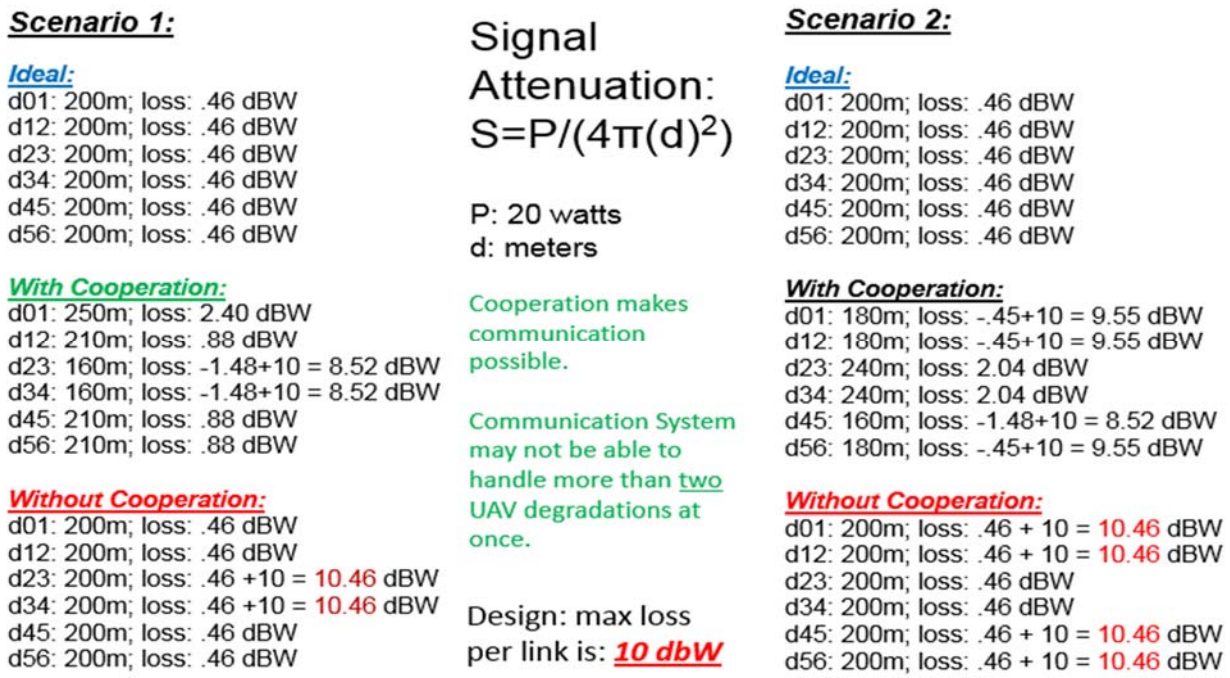

Figure 9. Loss in Decibels per Communication Link

The parameters and equations used in our experiment are:

- $S = P_t/(4\pi d^2)$, d is in meters
- Pt = 20W (13 dBW)
- N0 = 2W (3 dBW)
- BER requirement = 10^{-4}
- Bit Rate, Rb = 100,000 bits/sec, Throughput = 0.5
- SNR = S/N0 = 20W/2W = 10
- SNR = S/NO = 13 dBW/3 dBW = 10
- SNR = (Rb*Eb)/N0
- Eb = SNR*NO/Rb = 10*2/100,000 = 0.0002 = $2*10^{-4}$
- f = 1 MHz
- RCV Gain, Gr = 15 dBW
- XMT Gain, Gt = 3 dBW
- Power in dBW = 10*log10(Power/1W)
- Power in W = $10^{(\text{Power in dBW}/10)}$
- FSPL = 20*log10(d) + 20*log10(f) + 20*log10(4π/c)-Gt-Gr

CONCLUSION

The use of simulations for visualizing and understanding the effects and influence of system parameters on relay chains is valuable to a large variety of UAV communication applications. We have presented an efficient game theoretic algorithm for UAV relay positioning. We simulated the amount of time required to achieve a desired UAV positioning as a function of Q-Learning rate for two scenarios. System degradations were introduced into the simulation in order to quantify the operating boundaries for achieving quality communication.

An innovative, self-forming, self-organizing, cooperative, autonomous system of distributed UAV communication nodes can be realized wherein each UAV acts collectively and cooperatively. This will enable a multi-UAV network's communication links to be made more resilient, resulting in enhanced levels of network availability and improved service quality. We have simulated the guidance of vehicle positions based on decisions from game theory, with each vehicle repositioning to achieve optimal performance.

A cooperative reasoning engine was implemented to ensure that the entire system functions optimally, as determined by communication quality and system availability. We arrange UAV formation based on distance and required network throughput. Self-positioning of UAVs is a function of cooperation that ensures that struggling UAVs get help from neighboring UAVs in the form of shorter signal relay distances. This results in overall enhanced system performance, as intermittent signal issues are either prevented or resolved more quickly.

Our simulation was implemented in Matlab, which can be ported to other platforms to enhance processing speed, if desired, for more complex system modelling. The advantage of our simulation is the ability to handle non-linear input parameters. Adding more UAVs is also relatively easy since each vehicle is modeled as a one-sided game with system oversite harmonization for cooperation.

REFERENCES

Bletsas, Aggelos, et al. (2006). Cooperative Diversity with Opportunistic Relaying. *IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006. Vol. 2.*

Burdakov, Oleg, et al. "Relay positioning for unmanned aerial vehicle surveillance." The international journal of robotics research 29.8 (2010): 1069-1087.

Cohen, R., Rahmes, M., Fox, K., Lemieux, G. (2016). Q-Learning Multi-Objective Sequential Optimal Sensor Parameter Weights. *IMCIC*

Demers, A., Keshav, S., Shenker, S. (1989). Analysis and Simulation of a Fair Queueing Algorithm. *In Proc. of SIGCOMM, pages 1–12*.

Hado van Hasselt. (2012). Reinforcement Learning in Continuous State and Action Spaces. Reinforcement Learning: State of Art. *Springer, pages 207-251*.

Klein, Daniel J., et al. "On UAV routing protocols for sparse sensor data exfiltration." American Control Conference (ACC). IEEE, 2010.

Roy, O. (2010). Epistemic Logic and the Foundations of Decision and Game Theory. *Journal of the Indian Council of Philosophical Research*, *27*(2), pp. 283–314.

Simone Parisi, Matteo Pirotta, Nicola Smacchia, Luca Bascetta, and Marcello Restelli: (2014). Policy Gradient Approaches for Multi-Objective Sequential Decision Making. *IJCNN 2014: 2323-2330.*

Wayne Winston. (2003). Operations Research Applications and Algorithms *4th. Edition.*