Automomous Control of Aircraft for Communications and Electronic Warfare: Insights and Promises of Recent Literature

Sean Carver, Ph.D. at Data Machines Corporation

May 31, 2021

Abstract

We pose an unsolved problem in autonomous control of aircraft for communications and jamming (electronic warfare) and review the literature relevant to this problem. Some work offers approximately optimal solutions to related problems in different domains—promising applicability to the important scenarios considered here. Other work covers methods that we may find useful in extending these relevant solutions.

The problem we address lies within the fields of adversarial Multi-Agent Reinforcement Learning (MARL) and active sensing. In our problem, two opposing factions (labeled "blue" and "red") compete to win a zero-sum/purely adversarial game. The blue side tries to maintain communication links between ground-based assets with a fleet of "comms;" whereas the red side tries to jam this network with a fleet of "jammers." An Unmanned Aerial Vehicle (UAV) becomes a comm or a jammer when fitted for one of these purposes.

Each faction lacks knowledge and access to the state of the opposing side, but benefits from inferring this state probabilisticly through positioning its fleet for best sensory performance and localization (active sensing). This maneuvering should take into account the real possibility of any UAV getting shot down by its adversary's appropriately positioned ground troops. That said, the objective of each side should remain foremost. The blue side aims to simultaneously keep all units in communication while the red side aims to simultaneously jam this communication. Despite best efforts, different units/UAVs can fall in and out of communication with their respective headquarters, making each of the blue and red factions a multi-agent collection, fully cooperating among itself, despite different information, to fight its adversary having opposing goals. Our contribution poses this problem while pointing to literature for possible ideas for moving the field forward towards a successful implementation for the full adversarial problem in real-world combat.

1 Introduction

If unfortunate circumstances compel our leaders to order our armed forces to take a city from an adversary, the command headquarters on the ground would benefit from constant two-way communication with all its other units during the conflict.

In the fog that accompanies such struggles, our forces cannot rely on our enemy's network of cell towers to keep in touch. Instead, two way radios, linked by a network of "comms" (UAVs for communication) will hopefully allow our friendlies to stay connected.

While vastly better than cell phones, such a network has it own set of challenges. Indeed, our adversaries clearly prefer to keep us out of communication. To pursue this preference, they may send up jammers (UAVs for blocking communication). Thus begins a delicate dance of each side positioning its fleet to best find the other's birds and in so doing best keep or block communications.

We study the question of how each side can control its fleet by autonomously ordering and carrying out flight and communications- electronics operation instructions (CEOI) to optimally achieve its objectives. We care about strategies for both sides, because to defeat our enemy, we must understand the intelligent countermeasures they may take. Moreover, in a real war, our side—as well as theirs—may choose to fly both comms and jammers, requiring strategies for both roles.

Recent literature has tackled the problem of near-optimal search and rescue [1] and other related search and localization paradigms [2, 3, 4]. Other work has considered different applications requiring similar tools—notably cyber-security [5] and precision farming [6]. Search and rescue, for example, clearly relates to the problem at hand because, as with rescue, each side in our conflict benefits from successfully inferring the positions of targets on the other side. But electronic warfare differs from search and rescue. People needing rescue presumably want the search effort to succeed and will presumably not try to block attempts to determine their positions. In electronic warfare, on the other hand, targets aim to conceal their true locations from their adversaries. As a result, while search and rescue can succeed with a purely active sensing and optimal control solution, in our scenario, we need to learn to counter an opponent's strategy. To this end, we propose to apply artificial intelligence: specifically, adversarial multi-agent reinforcement learning. This paper reviews the literature relevant to this approach to victory.

2 Optimal and sub-optimal filtering

The filtering problem takes measurements of a stochastic system—possibly transformed measurements and possibly with noise—and produces estimates of the state of the system. This section reviews classical work on this problem, citing textbooks instead of recent papers.

Readers will find the optimal solution to this problem in the first pages of many textbooks on nonlinear filtering [7, 8, 9, 10]. The solution implements a recursion consisting of alternating applications of the Chapman-

Kolmogorov Equation and Bayes Rule.

Unfortunately, solving each of these equations demands an integration remaining provably intractable in most cases—indeed in all but two cases that researches have already identified. In all other cases, a researcher must settle for an approximation—a sub-optimal (but hopefully still *approximately* optimal) filter. Readers will find that the rest of the nonlinear filtering textbooks (the rest beyond the first few pages devoted to the optimal exposition) develop these sub-optimal approximations.

First, we list the two truly optimal solutions to the filtering equation as (1) the Kalman filter, and (2) the finite hidden Markov filter. The Kalman filter uses a linear model of the process, a quadratic objective function measuring optimality, and Gaussian noise corruption (LQG problem). The Gaussian assumption must hold for both the process noise and the measurement noise. The LQG problem divides into the linear quadratic estimator (LQE) problem for the optimal state estimates and the linear quadratic regulator (LQR) for the optimal control of such a system. As we will see below, the solutions of these problems decouple in a surprising and useful way (the separation principle, see the last section below) but unfortunately, this decoupling makes use of the specific assumptions we have imposed here.

Likewise, a finite hidden Markov filter uses a finite-state model of the process. These restrictions unacceptably constrain usable models for our application area, and therefore we will focus on approximately optimal alternatives.

Several specific approximations merit mention. An extended Kalman filter linearizes the state space around each sample point allowing the calculations behind the Kalman filter to proceed approximately, even when the assumptions for using a Kalman filter do not hold exactly. The approximation works well when the optimal probability distributions for state remain close to Gaussian. If they do not remain approximately Gaussian, the Extended Kalman Filter can perform poorly, leading to poor state estimates and impoverished inference.

A second approximation, a grid-based filter, approximates the state space with a finite grid of points allowing the calculations behind a hidden (finite) Markov filter to proceed, even for infinite state spaces. A grid-based filter works well for the lowest dimensional state spaces, but becomes computationally intractable when the dimensions become even slightly higher. In preliminary investigations the electronic warfare problem that motivated this review [Carver, research paper in preparation], one and two targets worked well (each adding two dimensions, longitude and latitude, to the state), whereas three simultaneous targets remained expensive beyond reach. In this work, we aimed to find jammers ("targets") without bearing information from observing successful or unsuccessful radio connections to friendlies.

The field calls the last class of filtering approximations that deserves our attention "particle filters." In short, the idea approximates evolving distributions with a finite swarm of Monte Carlo sample points called particles. These methods possess great generality and flexibility, but many researchers find particle methods more difficult to understand, and to successfully implement, than their simpler and more straightforward cousins.

Note that there exist many different ways to implement particle filters, each with its own benefits and limitations. We will discuss these methods further in the next section, as several papers concerning Active Sensing use particle filters.

3 Active sensing

Active sensing solves a control problem, and as such has considerable overlap with reinforcement learning. Both domains use observations to select actions with some notion of how to make that selection. Whereas reinforcement learning tries to optimize cumulative reward, many implementations of active sensing choose its action (they call it control) to maximize information or, equivalently, minimize entropy, in the distributions for the estimated quantities. For example, in active sensing for search and rescue, we want to control the sensors (choose the action) to best reduce the uncertainty in the target locations.

The electronic warfare application that we consider does not have exactly this form. We aim not just to locate the other side's birds, but also we aim to keep (or for red, block) communication between blue units. This additional objective lands the problem squarely in the purvue of reinforcement learning. That said, the objective of electronic warfare benefits from knowing, even with uncertainty, the locations of the other side's UAVs. Moreover, like with active sensing, each side can move its own fleet to localize the other's, but, with electronic warfare, it becomes a means to the end of meeting its other objectives. Therefore, we look to the literature on active sensing only for inspriation.

We start with one representative paper in this field [1], by Hoffman and coauthors. Hoffmann et al. considers the problem of target localization, with a search and rescue application in mind. They consider a number of mobile sensor vehicles (analogous to blue's comms) and a number of fixed-location targets (analogous to the red's jammers that blue tries to find by moving its comms). While Hoffmann uses fixed targets, he cites [4] for a motion model that would generalizes his solution to moving targets. Both of these papers advance particle filter solutions. Hoffmann specifically closes the loop with a control that minimizes the uncertainty in the target locations. In that sense, Hoffmann implements an active sensing paradigm.

It becomes interesting to consider the consequences of the change in objective function needed for our application. For search and rescue, the mobile sensors care only about the uncertainty in the state distribution of the targets to move its fleet to localize the targets. Our application changes this picture: if there exist directions of variation that do not matter as much for the objective of keeping targets in communication, our agents may care less about the uncertainty in these directions.

We propose to codify the comms and jammers objective function by giving a reward to each agent based on which units they succeed in directly or indirectly (eg. through secondary connections) contacting. Each agent then tries to maximize the cumululative discounted expected value of this reward.

Hoffmann considers two cases separately: bearings-only measurements, and range-only meausurements. Both will become useful simplifications in our case. Taking inspiration from Hoffmann, we can test the performance of our agents under each of these constraints (ie. bearings-only and range-only measurements) separately, then together in both homogeneous and heterogeneous mixtures of constraints across sensors.

Extending the work of Hoffmann, Ryan et al. [2] use non-trivial models for the sensor dynamics appropriate for fixed-wing UAVs that may tens of seconds to execute a maneuver, such as a 180 degree turn. In constrast, the quadrotor robots [11] of our models have simple dynamics and can essentially execute maneuvers in a single time step. While our work will initially consider UAVs of the sort that Hoffmann considers, Ryan's paper generalizes the problem statement in an important direction in that it removes an assumption about the sensor platforms that does hold for the quadrorotor devices we model here, but fails for other UAVs, such as fixed-wing aircraft.

Another paper [3], by some of the same authors, presents the related particle filter-based unifying approach to search and tracking.

4 Reinforcement learning and extensions

This section spans several disciplines, including reinforcement learning, deep reinforcement learning, distributional reinforcement learning, Bayesian reinforcement learning, and multi-agent reinforcement learning (see citations below).

We start each topic by defining the relevant terms listed above, and providing one or more citations. Reinforcement learning (RL) [12, 13] extends machine learning to sequential problems where an agent or agents learn to interact with an environment to maximize cumulative reward.

Researchers have solved similar problems in different disciplines. Indeed, RL shares a sizable overlap with control theory in engineering, including active sensing. In engineering, adaptive control [14, 15] lies closest to RL. The terminology of engineering differs from the terminology of reinforcement learning, but the terms map perfectly to one another. The "controller" (agent) interacts with the "plant" or "process" (the environment) by selecting a "control" (action) that "minimizes cost" (maximizes reward).

Engineers typically deal with continuous systems (ie robotics), whereas many, but not all, reinforcement learning solutions have finite time steps, finite action spaces, as well as, often, finite state spaces. The treaatment remains essentially the same, but may look a little different. For example, in engineering an analogous integral of cost can replace the RL return consisting of a discounted accumulation of rewards.

Engineers have fully developed the linear theory, which reinforcement learning has barely touched. In another difference, adaptive control, often aims more to maintain control of the plant in the face of natural but unpredictable changes, rather than, as aimed for often in RL, to learn to control the plant *de novo*.

Despite cosmetic differences, if both systems solve the same problem,

and to the extent that they solve it optimally—with the same optimality criteria—the solutions must tautologically coincide. But the approximations to optimality made in each discipline may differ, and the assumptions and problem formulations can also differ, as well. Undoubtedly, we would find a similar truth to analogous statements made of connections betweeen RL and active sensing.

Deep RL [16] uses neural networks to represent the functions learned by the agent(s). Classically, RL implementations, including many deep RL implementations, deal with inevitable uncertainty in represented quantities by maintaining best point estimates for these quantities. Distributional RL [17] departs from this tradition by maintaining full probability distributions for the uncertain quantities. If the actor(s) also perform Bayesian inference on these distributions (as they generally do), the agent implements Bayesian RL [18].

As an approach to our application, we would like to direct the reader's attention to distributional RL, its slightly smaller subset, Bayesian RL. Consider that the target localization and active sensing literature of interest represent uncertain quantities with distributions, just like distributional RL. While an RL algorithm could derive point state estimates from distributional state information to decide on an action, such an approach seems wasteful. Achieving an optimal solution to the electronic warfare problem stands as a worthy, if impossible, ambition for a machine learning engineer. That said, to not use the full distribution returned by filtering amounts to giving up on this ambition, or so it seems.

However, the mathematics can sometimes work out so that throwing away the distribution for a point estimate succeeds as the optimal solution [19]. For example, to control a linear regulator, where a quadratic cost function determines optimality, corrupted by Gaussian noise, (the LQR problem) the optimal solution uses a Kalman filter to produce the optimal state (point) estimate as a function of time. Then the solution applies the optimal control for a determinisitic regulator given that the "assumed known" state equals the state (point) estimate. In this case, throwing away the distribution allows a simpler, more parisimonious, and indeed still fully optimal solution.

A system satisfies the so-called *separation principle* if such a separation between estimation and control holds for the system. But systems do not always satisfy the separation principle—those that do stand as the exceptions. If an engineer can get away with invoking the separation principle, the calculations greatly simplify. Indeed, for the LQR problem, the solution exists in closed form.

We do not know how well the separation principle applies in our problem and we do not know how computationally expensive it will become to use purely distributional methods. The approach we suggest applies distributional RL (or, more precisely, Bayesian RL) to simplified versions of the problems first, then push the envelope both in terms of the complexity of the scenario, and in terms of the approximations used—such as the separation principle hoping that it applies at least approximately.

In the LQR problem we have a precise notion of optimality (the quadratic cost function) and a proof that the solution presented optimizes this condition. In our comms and jammers scenario we have nei-

ther. Moreover, a lengthy (and uncertain to succeed) search for such a proof may have little to no value for our clients.

We could say, however, that we will apply principles, such as Bayes Rule, that researchers in other contexts have shown lead to provably optimal solutions. In that sense we could say that "optimal" principles underlie our methods, rather than that our solutions achieve optimality. Approaching the problem in this way, we hope that our solutions will stand as very good, indeed good enough, and perhaps even the very best—unproved, of course.

Finally multi-agent RL [20] extends RL to environments that include other interacting agents cooperating or competing for reward. The definition of "multi-agent" given the introduction to Shoham's textbook [21] requires that the agents have different information, different interests (quantified as reward), or both.

By this definition, Hoffmann's implementation of search and rescue [1], discussed above, does *not* qualify as a multi-agent system because (1) all agents share all relevant information over a radio channel, and (2) they all have the same interests in localizing the target(s). This conclusion holds even though each sensor platform behaves independently of the others and performs its own calculations (which, because of shared information, the authors expect to remain identical to the calculations of the other agents, as the search proceeds).

On the other hand, our comms and jammers implementation does land squarely in the multi-agent camp because the blue and red sides certainly do have different interests, and moreover, barring unmodeled espionage, certainly do have different information. Indeed, the two sides do not just have different interests, they have diametrically opposed interests.

To reiterate, each faction has its own sensors with which it takes measurements to share among its team—but hide from its adversaries. But even within each faction, some units can access information that other units cannot. Indeed, the individual blue and red units may yet find themselves out of communication with their compatriots. Thus, even assuming factions share interests, they may not always share all information. As a result, there can exist more than two agents (by the definition Shoham quoted) in our environment. These realities become challenging wrinkles which we must deal with for success.

Reinforcement learning algorithms divide into *model-based* algorithms and *model-free* algorithms. The two methods differ in that a model-based algorithm maintains a model of the environment and uses this model to select its actions. Model-based reinforcement learning would fail catastrophically in our real-world scenario. The model in the model-based algorithm would have to include submodels of all agents, including those of our adversary. But we must expect that our adversaries will exploit any modeling assumptions we make—and posing a model of our adversaries requires making such assumptions.

The model-free alternative presents itself as an option. Within Bayesian reinforcement learning, a recent comprehensive survey of BRL [18] discusses and develops only two such classes of algorithms in that context: *Bayesian policy gradient algorithms* [22, 23] and *Bayesian actor-critic algorithms* [24, 23]. Of these two possibilities the authors of [18] caution

that their Bayesian actor critic implementation takes advantage of the Markov property, but that the Bayesian policy gradient algorithm does not make this assumption and stands appropriate for partially observed problems including Markov games.

These constraints, together with an initial preference for BRL, narrow, to a single option, the initially wide field of RL algorithms of interest along this path. However, there remain other trade offs to consider, so we should not focus on purely Bayesian methods. Indeed, it seems that just one set of authors has produced most of the work on model-free Bayesian RL, with the rest of the community slow to adopt these (admittedly somewhat recent) ideas. Moreover while these authors discuss applying their methods in a multi-agent framework, they intend their work mostly for single- or few-agents. They concede that their methods require a posterior over the policies of all agents. In some cases, the calculations may remain tractable, such if the agents have independent policies. But this simplification seems questionable for our application. The authors also mention the possibility of addressing this constraint with a myopic (1-step) look ahead on the value function and cite [25], an article about Q-learning. It remains unclear to what extent agents employing BRL will have comparable capabilities to the capabilities of agents trained with other methods—methods intended for, and tested on, primarily multi-agent environments. Nevertheless, we deem that these Bayesian techniques still merit a close look, careful understanding, and revealing tests against other methods.

Both Bayesian policy gradients and the Bayesian actor critic model the time varying gradient of the objective function (used to train the policy) as a Gaussian process (GP)—a type of stochastic process on the real line, or, in our case, a discrete subset where we have data. This objective function gives the expected return (discounted sum of rewards). To satisfy its definition, a GP must have all its finite dimensional distributions (ie. the joint distribution of a finite sample of its domain) possessing a multidimensional Gaussian distribution. With GPs, Bayesian policy gradients (and Baysian actor-critic) can perform inference on the posterior of the gradient conditioned on the data, which allows for the success of the method. Still it remains unclear to what extent the Gaussian process assumption has limitations.

For this reasons, together with the lack of testing on multi-agent systems, we cast a wider net by considering recent work of several authors concerning other techniques for multi-agent RL, particualarly those specifically intended for adversarial applications.

In a recent paper, Lowe et al. [26] considers an actor-critic paradigm, but one that imposes centralized training to prepare for decentralized execution. Lowe and coauthors cite a recent body of previous work briefly reviewed in their paper. In this paradigm, for all agents, there exists one single central critic trained with the all agents observations (or the whole state), and all actions (or all policies) of all agents. After training, during testing and execution no centralizations remain, and each agent executes its own decentralized but pre-trained critic. On the other hand, each agent's actors in this paradigm remains always decentralized. Though the actors and never get the benefit of centralized training, the success of Lowe's methods apparently depends upon consistently decentralized

actors.

A second paper by Srinivasan et al., [27], considers an alternative approach which never centralizes the training. Decentralized training, as with decentralized testing and execution, may matter to us if we want to train our agents as they perform in the field. Of course this property matters most in situations in which we must adapt and counter, on the fly, to our adversaries adapting and countering us. Indeed Scrinivasan and coauthors intend their work for adversarial

A different paper, [28], considers multi-agent actor-critics in situations where different agents cooperate to maintain a graph of two-way connections—a relevant capability for comms and jammers within each faction, though, in this paper, without the spector of competition from the other side.

Finally Foerster et al....

References

[1] G. M. Hoffmann and C. J. Tomlin, "Mobile sensor network control using mutual information methods and particle filters," *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 32–47, 2009.

[2] A. Ryan and J. K. Hedrick, "Particle filter based information-theoretic active sensing," *Robotics and Autonomous Systems*, vol. 58, no. 5, pp. 574–584, 2010.

[3] J. Tisdale, A. Ryan, Z. Kim, D. Tornqvist, and J. K. Hedrick, "A multiple UAV system for vision-based search and localization," in *2008 American Control Conference*, pp. 1985–1990, IEEE, 2008.

[4] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 425–437, 2002.

[5] D. Nicholson, S. D. Ramchurn, and A. Rogers, "Information-based control of decentralised sensor networks," in *Defense Industry Applications of Autonomous Agents and Multi-Agent Systems*, pp. 15–32, Springer, 2007.

[6] E. Testi, E. Favarelli, and A. Giorgetti, "Reinforcement learning for connected autonomous vehicle localization via UAVs," in *2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, pp. 13–17, IEEE, 2020.

[7] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house, 2003.

[8] D. Crisan and B. Rozovskii, *The Oxford handbook of nonlinear filtering*. Oxford University Press, 2011.

[9] A. Smith, *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[10] N. Chopin and O. Papaspiliopoulos, *An introduction to sequential Monte Carlo*. Springer, 2020.

[11] G. Hoffmann, H. Huang, S. Waslander, and C. Tomlin, "Quadrotor helicopter flight dynamics and control: Theory and experiment," in *AIAA guidance, navigation and control conference and exhibit*, p. 6461, 2007.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[14] K. J. Åström and B. Wittenmark, *Adaptive control*. Courier Corporation, 2013.

[15] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "Reinforcement learning and optimal adaptive control: An overview and implementation examples," *Annual reviews in control*, vol. 36, no. 1, pp. 42–59, 2012.

[16] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[17] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," *arXiv preprint arXiv:1806.03335*, 2018.

[18] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar, "Bayesian reinforcement learning: A survey," *arXiv preprint arXiv:1609.04436*, 2016.

[19] K. J. Åström, *Introduction to stochastic control theory*. Courier Corporation, 2012.

[20] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[21] Y. Shoham and K. Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[22] Y. Engel and M. Ghavamzadeh, "Bayesian policy gradient algorithms," *Advances in neural information processing systems*, vol. 19, p. 457, 2007.

[23] M. Ghavamzadeh, Y. Engel, and M. Valko, "Bayesian policy gradient and actor-critic algorithms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2319–2371, 2016.

[24] M. Ghavamzadeh and Y. Engel, "Bayesian actor-critic algorithms," in *Proceedings of the 24th international conference on Machine learning*, pp. 297–304, 2007.

[25] R. Dearden, N. Friedman, and S. Russell, "Bayesian q-learning," in *Aaai/iaai*, pp. 761–768, 1998.

[26] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *arXiv preprint arXiv:1706.02275*, 2017.

[27] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling, "Actor-critic policy optimization in partially observable multiagent environments," *arXiv preprint arXiv:1810.09026*, 2018.

[28] J. Su, S. Adams, and P. A. Beling, "Counterfactual multi-agent reinforcement learning with graph convolution communication," *arXiv preprint arXiv:2004.00470*, 2020.