ICAS-98-3,7,1

A98-31555

CONTROL OF A HIGH ENDURANCE UNMANNED AIR VEHICLE

John Wharington*, Israel Herszberg†

* *Research student, Wackett Aerospace Centre, Department of Aerospace Engineering, Royal Melbourne Institute of Technology (RMIT), GPO Box 2476V, Melbourne, Victoria 3001, Australia.*
† *Head of Department of Aerospace Engineering, RMIT.*

ABSTRACT

In the quest for very high endurance, designers of Unmanned Air Vehicles may need to develop control systems that extract energy from gust patterns such as thermals. This paper presents control systems for two stochastic, optimal control problems that arise from this concept of soaring UAV: the first involves centering the vehicle about a thermal, and the second involves the use of thermals and inter-thermal gusts. In both cases, reinforcement learning provides the basis for high level, direct adaptive control that drives a low level control system. The low level systems were constructed with the aid of a heuristic and open-loop solution respectively. Simulation results indicate the success of the designs, though faster convergence is necessary for practical application.

INTRODUCTION

With endurance becoming a critical mission objective of Unmanned Air Vehicles (UAVs), automatic control will be an integral part of advanced designs. Autonomous control allows deployment of many vehicles by a relatively small support crew.

Range and endurance can be extended considerably by supplementing the on-board fuel supply with energy extracted from the environment. Solar power has been identified as a potential way to do this. The AeroVironment Pathfinder demonstrated the practicality of solar power in high altitude craft[1] and studies show that endurance of months is possible in summer near the poles.

However with current technology, endurance generally is limited to about a week, which is similar to that achievable with non-renewable fuels. Moreover, solar aircraft need very low wing loadings (6 kg/m^2) which tends to produce large, slow aircraft.

This restriction can be lifted by having the aircraft *soar*, by which the aircraft maneuvers so as to draw energy from various gust patterns. Factoring this into the design affords higher wing loadings and reduced battery size, enabling greater payload and higher cruise speed. Although vehicles employing this concept would need to operate in the troposphere, indefinite endurance may be possible in favorable weather.

Automatic control of soaring vehicles is challenging, especially given the severe limitations on available weight and power for avionics. Soaring is characterized by nonlinear dynamics and involves estimation of weather conditions. Not only should the control be stable, it should also be optimal in that the vehicle safety and mission performance depends on the vehicle's efficiency at both drawing energy from the wind and using its energy for transport at the greatest speed.

This paper reports on two control problems associated with this vehicle concept. The first is the fundamental problem of how to locate a thermal (rising column of air) and position the vehicle around its center. The second problem is how to use thermals and inter-thermal gusts for greatest average speed. Sensor uncertainty plays a significant role in both problems.

The authors address these problems with reinforcement learning, a relatively new method of direct adaptive control. This approach was pursued to investigate the feasibility of designing autopilots that improve over time and can adapt to changing vehicle dynamics or environment. Simulation experiments were performed on an IBM 686-PR200+.

FLIGHT SKILLS VIA REINFORCEMENT LEARNING

Reinforcement learning[2] describes numerical methods, that can be applied on-line, for learning an optimal control policy entirely from scalar performance feedback (i.e. a reward signal). It amounts to direct adaptive control, in that no system model is required. Not providing such *a priori* information comes at a price — convergence can be slow.

21st ICAS Congress
13-18 September 1998
Melbourne, Australia

A particular strength of this approach is the ability to optimize the control of systems with relatively arbitrary representations of states and actions. It can thus be used in learning to control the mode of a low-level controller or coordinate low-level controllers[3]. This capability can be used to overcome concerns of suboptimal performance, by designing a subordinate controller with an assured properties. It can also greatly speed convergence by simplifying the high-level task to be learnt.

Some roles of adaptive flight control (particularly of certain low-level systems) are unlikely to warrant the reinforcement learning approach. At present, it requires considerable exposure to the system during convergence and so cannot really be considered as a means of learning to compensate for rapid structural failure. When the plant can be modeled (and estimated) accurately and conveniently, the model can usually be used to derive a control system that adapts faster than would reinforcement learning.

Reinforcement learning may be seen as a technology to be adopted as a last resort, where the plant or objectives are such that mathematical treatments fail. However, it also deserves consideration because of the convenience it offers, even for use off-line.

Its methods are generally directed at solving Markov decision tasks, where the dynamic system evolves according to transition probabilities that depend on the state and control. Even though the presence of hidden states pose similar problems to reinforcement learning, as they do to other control systems, there is a certain capacity to learn strategies that are robust to uncertainty and perceptual aliasing. This is particularly advantageous in the context of weather-sensitive aircraft.

Several reinforcement learning control systems are available[4] and the basis of most of these can be traced to dynamic programming and the calculus of variations. These systems, favored by the engineering community for their mature foundations and transparency to analysis, are referred to as neuro-dynamic programming ('neuro' reflecting the use of neural networks for function approximation).

Q-learning[5] with CMAC neural networks[6] was used in the experiments described in this paper.

THERMAL CENTERING

Task description

The objective is to find the path that terminates in a circling maneuver about the thermal center and which maximizes the *net* lift over the course. Optimal paths implied by this objective are not minimum distance paths to a tangent to the thermal, since the net lift is a function of bank angle via its effect on vehicle sink. A balance must be established in the trade-off between tracking deviation and vehicle sink.

The following assumptions can be made to simplify the problem:

- The thermal is constant over altitude, reducing the task to two dimensions.
- Thermals drift at the same speed as the prevailing wind, so lateral wind can be disregarded.
- The vehicle speed is constant: for our purposes the maneuver starts with the aircraft at its safe thermalling speed.
- A lateral autopilot is in place to ensure turns are coordinated (no sideslip). By further assuming the autopilot is ideal (no delay or error), the task is focused on the higher level problem involving bank control only.

Implicit in the objectives is the determination of the optimum bank angle for circling once the thermal has been centered. As the maneuver terminates in this circling, the thermal profile needs to be estimated in advance.

The vehicle sensors comprise a navigation system and a total energy variometer, which measures vertical gust velocity at the aircraft and is subject to noise. In practice, variometers are also subject to lag, but this is neglected here.

The navigation system needs to measure heading, as a bare minimum, as well as position. This could be accomplished with a compass and an inertial navigation system or by two Global Positioning System units separated from each other.

Additional sensors may be required by the low-level lateral autopilot.

Heuristic methods

The problem can be decomposed generally into two elements according to a sensorimotor view (sensing plus control). First, a mechanism is required to translate the sensor measurements into information regarding the thermal such as position and profile. This is called the

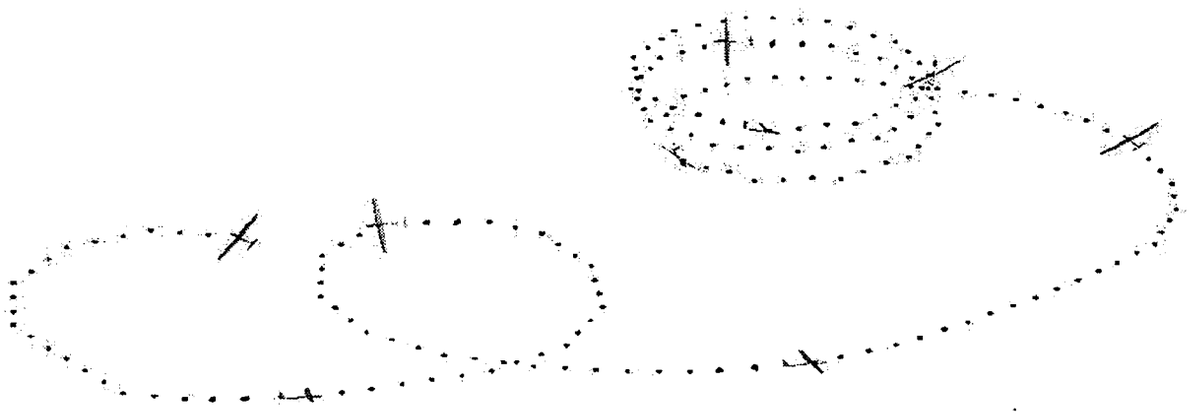

Figure 1: Demonstration of centering heuristic

thermal locator. Another mechanism uses this information to command the bank angle so as to move the vehicle closer to the thermal.

A rough cut at the problem is to implement a broad strategy employed by glider pilots. The strategies of expert pilots is probably reflexive, and it is difficult to extract these implicit skills. However, novices are taught with the aid of a simple heuristic, of which several have been proposed of varying efficiency and simplicity.

The heuristic proposed by Reichmann[7] (his method III) is a suitable starting point. Turns are widened when approaching the thermal and tightened when moving away, thus:

$$\phi = k_1 \left(1 - \cos(\theta)\right) + k_2 \qquad (1)$$

where ϕ is the bank angle, θ is the bearing to the thermal center, and k_1, k_2 regulate the variation of bank.

Here θ is the quantity to be produced by the sensor mechanism.

For such a simple heuristic, Equation 1 is quite effective. Figure 1 shows it in action.

Nevertheless, this rough method is not actually optimal with respect to our objectives. Clearly it needs to be extended to account for the thermal profile and variometer errors.

Optimal centering

In practice, Reichman's heuristic as presented above is oversimplified. Pilots implicitly estimate thermal profile as well as relative location, and effectively refine the gains of the above scheme as dictated by optimality and uncertainty. In particular, these gains may vary with the distance to the center.

The numerical determination of the optimal gain schedule can be cast as a reinforcement learning problem where the net lift (as measured) acts as a reward signal. This could be used on-line to adjust for slow variations in the vehicle aerodynamics, or deteriorating errors in the variometer.

Greater control over the thermalling heuristic is made possible by the replacement of $\cos(\theta)$ in Equation 1 by the sag-cosine function defined as

$$\begin{aligned} C_\pi &= \frac{e^{-a_3}}{e^{|a_3|}} \\ C_\theta &= \cos(\theta)\frac{e^{a_3\cos(\theta)}}{e^{|a_3|}} \\ \text{sagcos}(\theta) &= \left(\frac{C_\theta - 1}{C_\pi + 1} + 1\right) \end{aligned} \qquad (2)$$

The heuristic, once normalized, becomes

$$\phi = \frac{\pi}{2}\left[a_2(1 - a_1)\,\text{sagcos}(\theta) + a_1\right] \qquad (3)$$

The parameters have been transformed to produce valid heuristics in the range $|a_1| \leq 1$, $|a_2| \leq 1$. The parameter a_3 regulates the degree of sag, as illustrated in Figure 2.

Trajectories are associated with an information value as well as the performance (reward) payoff. The information value is related to the exploration of the environment that is required by the thermal locator. To illustrate this, consider the functioning of the thermal locator when the path segment is linear (or nearly so) in the early stage of search (Figure 3). The location problem is under-determined and there will be two potential locations of the thermal. Similarly, estimation of the thermal profile can also be affected by the vehicle path, resulting in range errors.

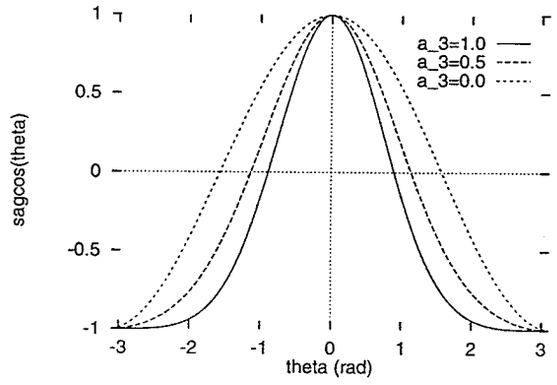

Figure 2: Sag cosine function

For completeness, the thermal locator needs to produce relative bearing, and thermal profile. A neural network will be constructed for this purpose.

The thermal locator comprises an array of nodes which perform regression on variometer measurements. Each node estimates the thermal profile assuming the location is at the nodes uniquely assigned location. A competition for the best fit establishes the resulting certainty and thermal profile and location returned by the system. This produces relative bearing, range, thermal radius and magnitude.

The model of thermal profile used in the regression is the exponential Gaussian:

$$w_T(D) = We^{-(D/R)^2} \tag{4}$$

where W and R is the vertical movement and characteristic radius of the core, and D the (lateral) distance to the thermal center.

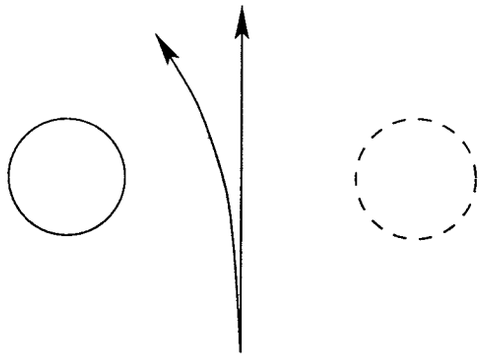

Figure 3: Underdetermination of thermal center. The dashed circle represents the illusory alternate location when the path is linear.

Demonstration

Before testing the system under reinforcement learning, it is helpful to optimize the system under ideal conditions (where ideal means a perfect thermal locator is in place). The benefits of doing this are twofold: useful ranges of the controls can be obtained, thereby reducing the search space when learning; and the ideal performance of the system is determined, enabling assessment of the learning system during convergence. The latter point is important because knowledge of some baseline performance, however idealized, can greatly assist in trouble-shooting and validation of the learning system.

In order to concentrate equally on weak or smaller thermals as strong, larger ones, the reward signal, $\overline{r}$ is defined as the net lift (updraft less sink) divided by the thermal strength W. The value $\overline{w}$ is a measure of the average reward over the trial period.

Centering behavior was considered from the initial encounter of a thermal, so that at the starting position of the aircraft, the updraft velocity equals a threshold value (0.15 m/s). Thermal size and strength were selected randomly in the range $R \in (40, 80)$ m, $W \in (3, 9)$ m/s.

The initial bearing of the vehicle was randomly selected in the thermal's hemisphere, as appropriate to the initial encounter case. The vehicle speed was held constant at 20 m/s, and a maximum load factor of 4 was imposed (limiting the maximum bank angle). The sink model had the form

$$w = -\frac{1}{LD}\left[\left(\frac{V_{LD}}{V}\right)^2 + \left(\frac{nV}{V_{LD}}\right)^2\right] \tag{5}$$

with parameters: minimum drag speed $V_{LD} = 25$ m/s, lift to drag ratio $LD = 35$. The resulting model expressed in terms of bank angle is

$$w = -0.286\left(0.640 + \frac{1}{0.640\cos^2(\phi)}\right) \tag{6}$$

This optimization was performed using simulated annealing[8], where the payoff (negative cost) was the sum of average reward over fifty trials, each lasting 100 seconds. Multiple trials were necessary so that the annealing process would find the best parameters despite the cost function being stochastic (due to the random initial heading). The importance-based sampling property of simulated annealing then promotes parameters that give the best overall performance, with a certain economy on the number of trials performed per cost evaluation.

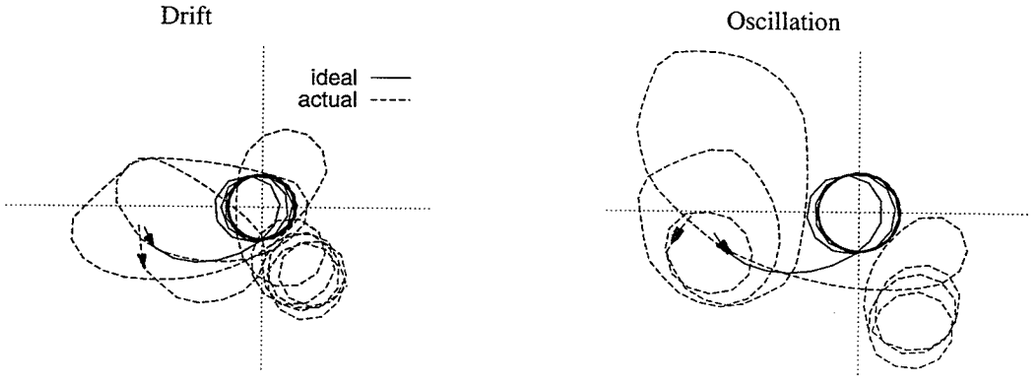

Figure 4: Plan view of trajectories showing failure modes of heuristic with scheduled optimal parameters.

Item	max	min
a_1	-0.771	-0.568
a_2	0.386	0.670
a_3	0.159	1.059

Table 1: Ranges of optimal parameters of thermalling heuristic under ideal conditions

Optimization under these conditions was performed separately for a range of thermals. The ranges of the resulting parameters are shown in Table 1. The first two parameters have relatively clear and uniform variation, so it is reasonable to fix the third parameter to a constant value when learning.

It is worthwhile to examine the performance of the system where these optimal parameters are scheduled according to the thermal profile as perceived by the neural locator system. The thermal locator comprised an 11 by 11 grid of units spaced 20 meters apart. Variometer noise was normal distribution of 0.3 m/s standard deviation. Measured updrafts less than 0.1 m/s were rejected.

The normalized payoff, averaged over the range of thermals was found to be 0.485 for the ideal case and 0.023 for the scheduled system. This indicates failure of the scheduled system owing to the disruptive effect of the thermal locator on the centering heuristic. Several distinct failure modes were identified in the trajectories of the scheduled system. These are illustrated in Figure 4 along with their ideal counterparts.

Clearly then, there is scope for improvement by reinforcement learning. A Q-learning system was set up to control the heuristic parameters a_1, a_2 within the ranges established above. The third parameter was set at 0.61.

For problems of this sort, it would be useful for the history of variables to be included in the state variables,

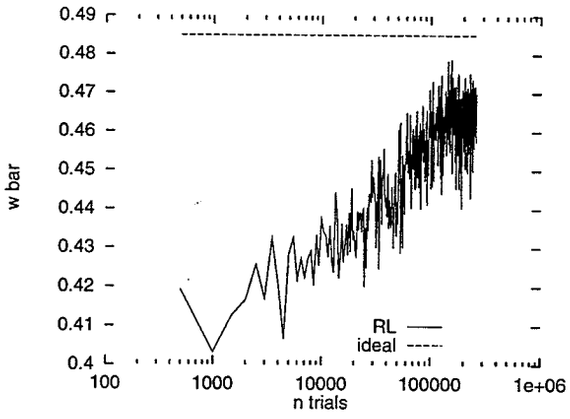

Figure 5: Thermal centering: Convergence history

so that the system could learn to react to *changing* data. However, for the sake of simplicity this avenue was not explored in this study.

The states of the reinforcement learning system were the thermal locator's estimate of the thermal size and strength, and the certainty of the winning unit.

Results of training are shown in the form of convergence history, in Figure 5, and a sample trajectory in Figure 6. Training was stopped once the normalized payoff reached within 2% of the idealized performance.

These results demonstrate the feasibility of tuning the centering heuristic via reinforcement learning, and suggests that reinforcement learning was able to find strategies that were robust with respect to the imprecision of the thermal locator. More detailed analysis is necessary to determine if, and how, these strategies actively hunt the thermal.

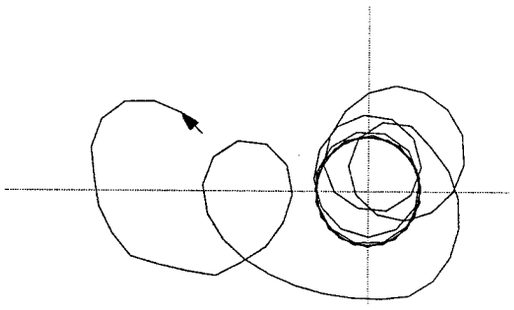

Figure 6: Robust centering of heuristic scheduled by adaptive control

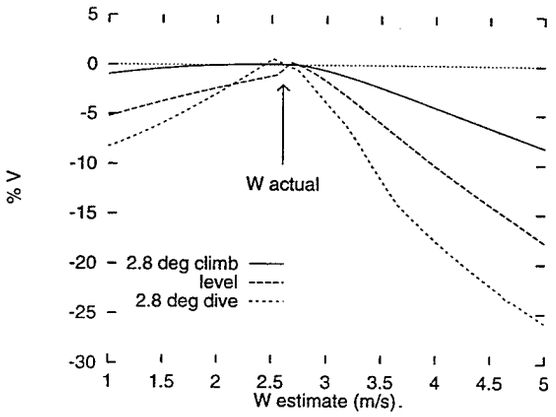

Figure 7: Speed loss due to estimation error in gust strength

Discussion

Convergence took eight hours of computation, largely due to simulation of the locator system. It would require several months in real-time, and so modest improvement is necessary to enable practical application.

Use of the heuristic as a low-level controller was critical to the success of the experiment. An attempt at optimizing the function $\phi(\theta, \ldots)$ directly may fail with current reinforcement learning algorithms. If the gains are subject to change over θ, the centering heuristic can be disrupted, slowing convergence as the task becomes much more complex.

The optimal path to the thermal center could be found by using the bang-bang type controllers derived by the calculus of variations. Optimal controllers of this type have been obtained for wheeled vehicles, e.g. [9]. Similar controllers apply here, since flight in the horizontal plane is equivalent to wheeled vehicles with bank angle analogous to steering angle. However, this approach would only be appropriate if the start position was distant from the thermal center, and would only be possible if the thermal location and profile were known in advance.

SPEED-TO-FLY UNDER UNCERTAINTY

Speed-to-fly theory is a method of determining how fast to fly in cruise between successive periods of circling in thermals. Metzger and Hedrick[10] developed a unified speed-to-fly theory that includes thermalling and dolphin flight (speed changes during cruise).

This theory is derived by the calculus of variations and considers the vehicle passage across a segment of known distribution of gust velocity. It determines the ring setting which singly dictates the strategy to be followed so as to reach a commanded height (or achieve a commanded glideslope) in minimum time.

The ring setting governs the speed of the vehicle in dolphin flight cruise as a function of gust strength. Following the dolphin segment, the glider climbs in a thermal (where possible) until the commanded height is reached. The relative activity of the two modes depends on gust strength and distribution.

Influence of gust estimation

In practice, the gust distribution is not known but must be estimated on the basis of past variometer measurements and perhaps aided by visual cues such as terrain and cloud features. In this study, we simplify the sensing issues by considering only errors in magnitude, though the distribution of updraft and sink would probably be more prone to error in practice.

Although it is possible to adjust the open loop policy as the estimate is improved, it is difficult to quantify just how this might be done. It is assumed here that one commits to the original speed ring setting, so the only adjustment that can be made during its execution is the time spent thermalling.

Errors in the gust estimate lead to height errors and speed losses. Figure 7 shows the resulting variation of speed loss to estimation error for climb, level flight and diving. Note that overestimates are more detrimental and that the loss characteristics vary with command glideslope.

These were calculated for a sailplane with minimum drag speed 25 m/s, lift to drag ratio of 49 and a maximum speed of 46 m/s. The gust distribution was a symmetric step.

Height errors that are incurred can be taken care of, by commanding the height proportional to the height error (as recommended by [11]). This will ensure that the height is bounded, but it isn't necessarily the optimal way of doing so.

Optimal height correction

Because average speed varies strongly and nonlinearly with gust strength, estimation error and commanded glideslope, there is scope for improvement to this system in several areas. Greater speeds would result, for example when the aircraft is too low, by waiting for a strong gust in which to climb.

The speed error curves of Figure 7 suggest that the estimator should be biased to underestimate in order to minimize the impact of error on speed. A similar effect might be achieved by adjusting the speed ring setting from that corresponding to the estimation. Indeed a common technique of glider pilots is to set the speed ring somewhat conservatively.

Attention is restricted to adjusting the height command system so as to improve the average speed and minimize height errors. This was cast as a reinforcement learning problem, where the states of the system are the estimated gust strength and height error. The control variable a adjusts the height command system:

$$h_c = -h + a \tag{7}$$

Reward signals were proportional to average speed less the square of the height error.

It should be mentioned that a separate reinforcement learning controller, in which the control variable was height command, was able to find an appropriate height correcting strategy, but did not sufficiently concentrate effort on optimizing speed.

The test environment consisted of an *independent* normal distribution of gust strengths of mean 2.54 m/s (500 ft/min) and standard deviation 0.568 m/s. Each gust segment was about 300 meters long. The estimator was modeled by adding noise to the actual gust strength and was of normal distribution of standard deviation 0.227 m/s. Independence means the task has the Markov property, though this restriction can be lifted with special mechanisms to address hidden states[12].

To establish baseline performance, the system was evaluated with commanded height $h_c = -h$, producing an average speed of 18.0 m/s and height mean square error of 24 m. Performance of the learning system during convergence, expressed as percent improvement over the baseline speed, is shown in Figure 8. The 4.5 percent speed improvement achieved is significant. This speed is also 4 percent faster than the baseline system with no estimation error. Height correction was improved by 8 %.

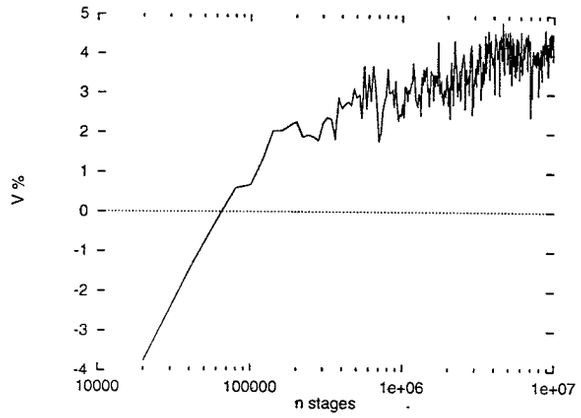

Figure 8: Convergence of speed-to-fly/height correction system showing improvement over baseline performance

Discussion

This problem was investigated largely as an exercise in adapting reinforcement learning to flight control. In addition, we were interested in quickly assessing the impact of uncertainty on speed-to-fly.

An optimal controller for this problem perhaps could be derived using an analytical technique, but this would require knowledge of various quantities, such as probability distributions, that may not practically be available to the designer. The strength of the approach adopted here is such quantities could be obtained by the learning system itself, while in operation.

For this problem, reinforcement learning was quite convenient and fast to simulate; time taken for the system to converge was only twenty minutes. However, convergence would have taken roughly six months of continual flight.

The learning system was successful in discovering, largely by itself, a control law that performed quite well despite significant sensor uncertainty. It would require substantial extension to be practical for on-line implementation, particularly because the real-time cost of the learning was high. Moreover, the system was adapted for *a specific environment.*

By building upon studies of this sort, we might be able to design more practical controllers (whether adaptive

or not) that function well on a whole range of environments. A support for such optimism is the potential for *generalized* control laws. Transformation of sensor and control variables and the use of neural networks may provide great assistance.

CONCLUSIONS

The control problems addressed in this paper contribute to an enabling technology for a novel concept of high endurance UAV. The results demonstrate the convenience offered by the reinforcement learning approach to adaptive control. However practical application awaits advances in convergence rate and generalization.

The experiments showed that the effects of uncertainty could be addressed appropriately by the control designs: in the thermal centering task, the controller helped 'reify' the unobservable thermal location and profile; and the controller for the speed-to-fly task both compensated for estimation error and exploited random variation in the environment. No attempt was made to find the system parameters for fastest convergence or best final solution.

Formulating the adaptive controller above a tailored, hard-wired controller was necessary to simplify the learning task. A secondary benefit of this is to provide a degree of protection against catastrophic failure.

It is hoped that this study will stimulate research into applications of reinforcement learning to flight control, particularly in autopilots that learn to improve fuel economy over the life of the vehicle.

REFERENCES

(1) N.J. Colella and G.S. Wenneker. Pathfinder: Developing a solar rechargeable aircraft. *IEEE Potentials*, pages 18–23, February/March 1996.

(2) Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.

(3) M. Huber and R.A. Grupen. Learning to coordinate controllers: Reinforcement learning on a control basis. In *15th International Joint Conference on Artificial Intelligence*, pages 1366–1371, August 1997.

(4) Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, (4):237–285, May 1996.

(5) C. Watkins. *Learning from delayed rewards*. PhD thesis, Cambridge University, 1989.

(6) W. T. III Miller, F. A. Glanz, and L. G. III Kraft. CMAC: An associative neural network alternative to backpropagation. *Proceedings of the IEEE*, 78(10), October 1990.

(7) H. Reichmann. *Cross-Country Soaring*. Thomson Publications, 1988.

(8) S. Kirkpatrick, C.D. Gelatt Jnr, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

(9) David B. Reister and Suzanne M. Lenhart. Time-optimal paths for high-speed maneuvering. *International Journal of Robotics Research*, 14(2):184–194, 1995.

(10) D.E. Metzger and J.K. Hedrick. Optimal flight paths for soaring flight. In *AIAA/MIT/SSA 2nd International Symposium on the Technology and Science of Low Speed and Motorless Flight*. AIAA, 1974. AIAA paper 74-1001.

(11) Risto Arho. Some notes on soaring flight optimization theory. *Technical Soaring*, IV(2):27–30, 1975.

(12) S.D. Whitehead and L.-J. Lin. Reinforcement learning of non-markov decision processes. *Artificial Intelligence*, 73:271–306, 1995.