Autonomous Soaring Using Reinforcement Learning for Trajectory Generation

Tim Woodbury, Caroline Dunn, John Valasek

January 15, 2014

vscl compos volatus

AEROSPACE ENGINEERING
TEXAS A&M UNIVERSITY

Outline

- Introduction
- Control policy
- Aircraft modeling and simulation
- Numerical results
- Conclusions

Autonomous soaring

- **Introduction**
- **Small electric unmanned aerial systems (UAS)**
  - **Capable sensing**
  - **Limited range/endurance**
- **Thermalling**
- **Soaring electric UAS**
  - **Comparable endurance to solar-powered4**
  - **Payload/performance flexibility**

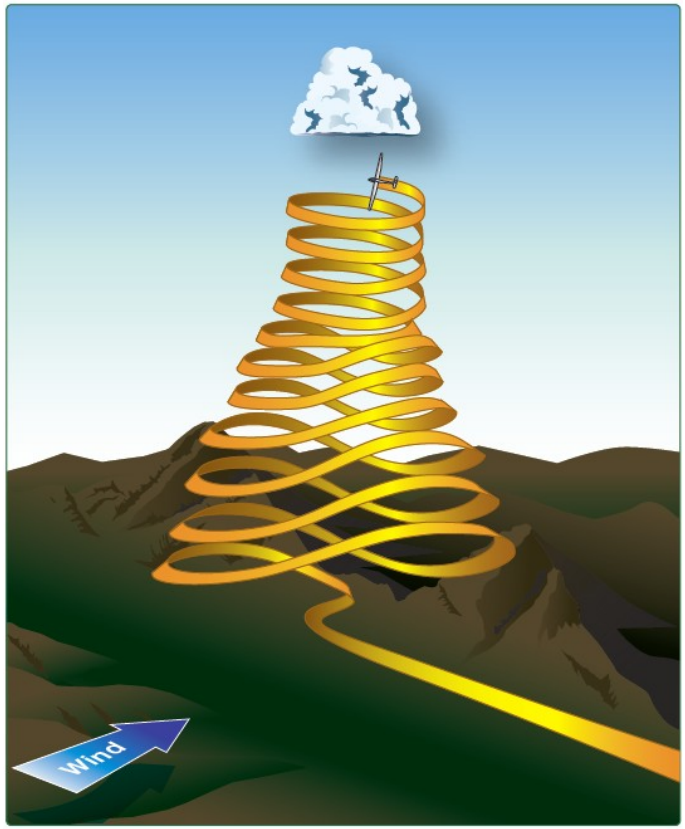


Image: *Glider Flying Handbook,* Chapter 10: Soaring techniques, Flight Standards Service, Federal Aviation Administration, 2013, FAA-H-8083-13A.

Literature review (Autonomous soaring)

Introduction

- Allen and Lin (2005,2007) estimate a thermal location based on an assumed model and energy time rate. Flight test results showed an average altitude gain of 173 m/thermal11,5
- Edwards (2008) presents flight test results; local updraft speed estimated and grid of nodes evaluated to determine the thermal center6
- Depenbusch and Langelaan (2010) implement receding horizon control to optimize energy gain given local knowledge of a wind field10
- Lawrance and Sukkarieh (2011) simultaneously explore and exploit a wind field using Gaussian process model for mapping with energy-efficient path planning 9
- Andersson (2012) et al implement stable thermal centering with flight test results using energy time rate for thermal identification4

Literature review (Thermal localization)

Introduction

- Most flight test implementations detect or estimate thermal locations by measuring aircraft energy and looking for external wind4,5,6
- Remote detection using visual or IR camera
  - Akhtar7 conducted field trials with an IR camera
    - Detected “hot spots” on ground and clouds
    - Could be used to infer thermal locations
  - Sheng et al.8 describe calibration of a thermal IR camera for UAS flights
    - Not intended for thermal soaring
    - Same architecture might be used for remote thermal detection

Introduction

- Autonomous navigation to thermal
- Novel approach to soaring: generate reference bank angle commands to reach and circle the thermal using RL
- Three advantages:
  - Generation of reference trajectories online reduced to table lookup
  - Robustness to model uncertainty
  - Q-learning gives flexibility to tailor performance via reward shaping

Introduction

- Sequential decision making based on experience interacting w/ environment
- Attempts to maximize rewards received
  - Given state *s*, select an action *a* (policy)
  - Based on the next state, receive a reward *r*
- Similar problem structure to dynamic programming
  - No knowledge of state transition function or reward structure req'd
- Constructs or approximates value functions:
  - State-value function *V(s)*
  - State-action value function *Q(s,a)*

Q-learning algorithm

Control policy

1. $Q(s, a)$ is initialized arbitrarily
2. For each episode in the learning:
   (a) Initialize at state s
   (b) For each step in the episode:
      i. Choose action a from s via policy
      ii. Observe the next state, s'
      iii. $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
      iv. $s \leftarrow s'$
   (c) Break when s is a terminal state

- α: step-size parameter; γ: discount-rate parameter; r : reward received

Q-learning for autonomous soaring (1)

Control policy

- Two assumptions
  - UAS has low-level feedback control
  - Thermal location is known
- Dynamic sailplane model implemented with Q-learning for lateral/directional guidance to thermal
- Q-learning: discrete-time commands at slow rate (~1 Hz)
- Continuous-time roll and pitch controllers
  - Roll control is proportional-integral-derivative (PID)
  - Pitch controlled by state feedback regulator
- Near thermal, different controller used for circling

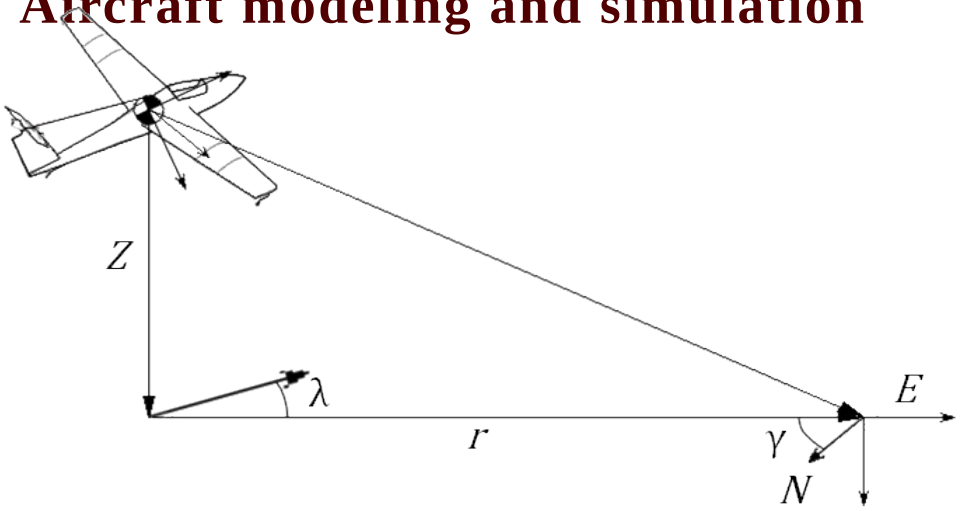
Q-learning for autonomous soaring (2)

Control policy

- Q-learning guidance algorithm:
  - Initialize the state: r (range to thermal), λ (azimuth to thermal relative to aircraft heading), φ (bank angle)
  - Loop over 500 discrete time steps, or until a break condition is met:
    - Choose action from available changes in bank: $\Delta = \{0, -\Delta\varphi, \Delta\varphi\}$
    - Simulate with commanded bank angle $\Delta+\varphi$ for TQ seconds
    - If $r > R1$, the maximum allowed range, receive a penalty reward and break.
    - If $r < R2$, the effective radius of the updraft, receive a goal reward and break.
    - Else, receive no reward and continue.

Q-learning settings

Aircraft modeling and simulation



- R1 = 1200 m
- R2 = 460 m
- |φ| < 30o
- TQ = 5 sec
- Δφ = 5o
- Δr = 50 m
- Δλ = 5o

- PID, LQR continuous-time control laws
- Updraft model (Allen 2005)
- Thermal size arbitrary

- Updraft radius steady level turns
- Updraft strength circling simulation length

Aircraft model

Aircraft modeling and simulation

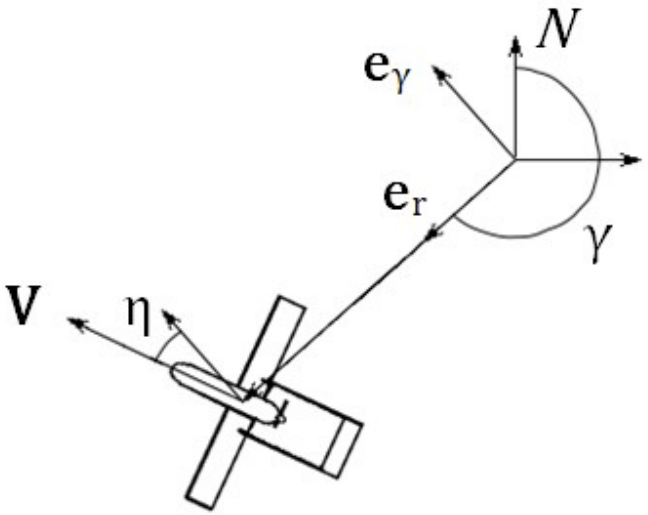
Linear decoupled longitudinal and lateral/directional models from Refs. [1] and [15]

Glider dynamics representative of small UAS

Schweizer SGS 1-36

Aircraft modeling and simulation

- Feedback control used to create baseline results of energy gains from circling thermal
- Future RL-based circling
- Lyapunov-based feedback control law designed
  - Guaranteed convergence* to desired circling radius
  - Feedback gain *KL* = 0.25 used
- Using Allen (2005) as a baseline, target radius is 0.6R2



$$g \tan \phi = r\dot{\gamma}^2 - \frac{U_1^2(\dot{\eta}_{ref} - K_L e_\eta) - \dot{r}^2\dot{\gamma}}{r\dot{\gamma}}$$

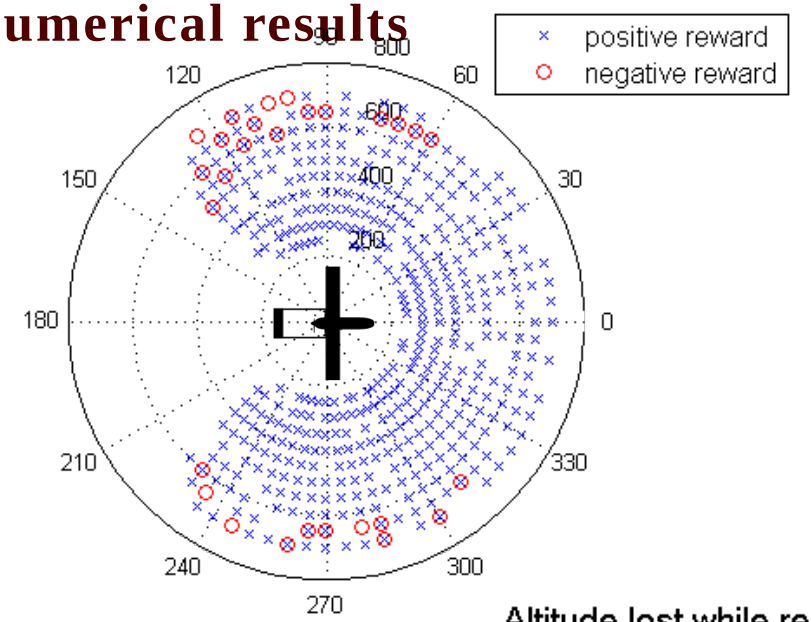
Simulation framework

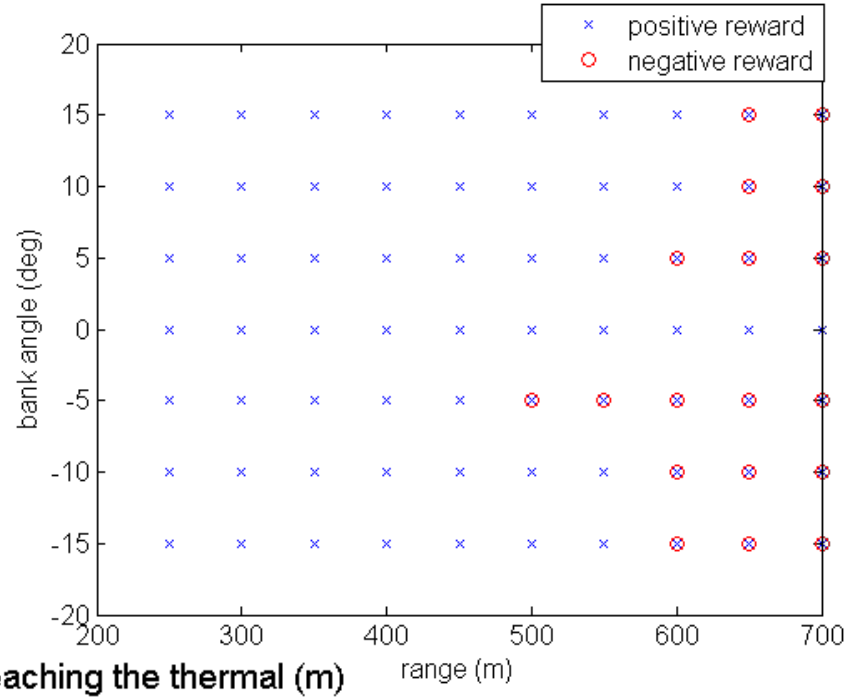
Aircraft modeling and simulation

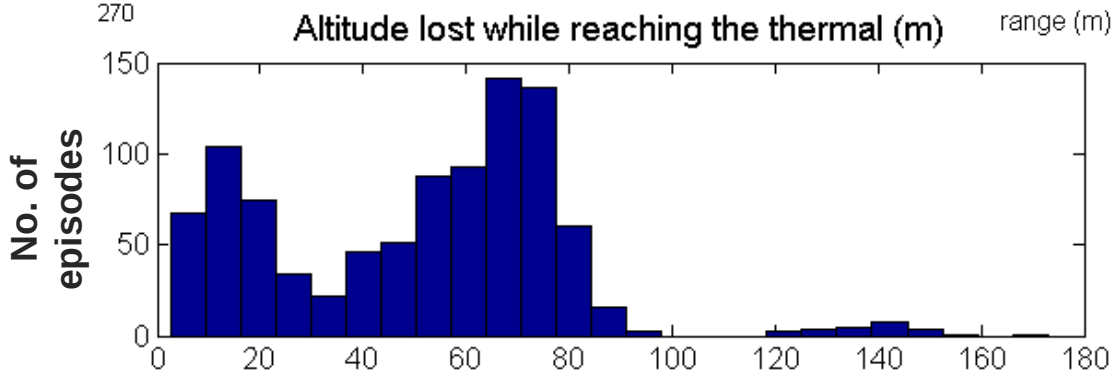
- RL navigation agent trained on 100,000 episodes, ε-greedy policy
- For learning Q, the state is initialized randomly within
  - 500 m ≤ r ≤ 700 m
  - -135o ≤ λ ≤ 135o
  - Φ = 0
  - Z = -300 m
  - All other perturbed states zero
- Evaluated w/1000 Monte Carlo simulations
- If aircraft reaches the thermal, circling controller used for 1000 sec
- Expanded initial conditions: 250 m ≤ r ≤ 700 m, -135o ≤ λ ≤ 135o, -15o ≤ Φ ≤ 15o
- 96.2% success in reaching thermal

Monte Carlo results (1)

Numerical results

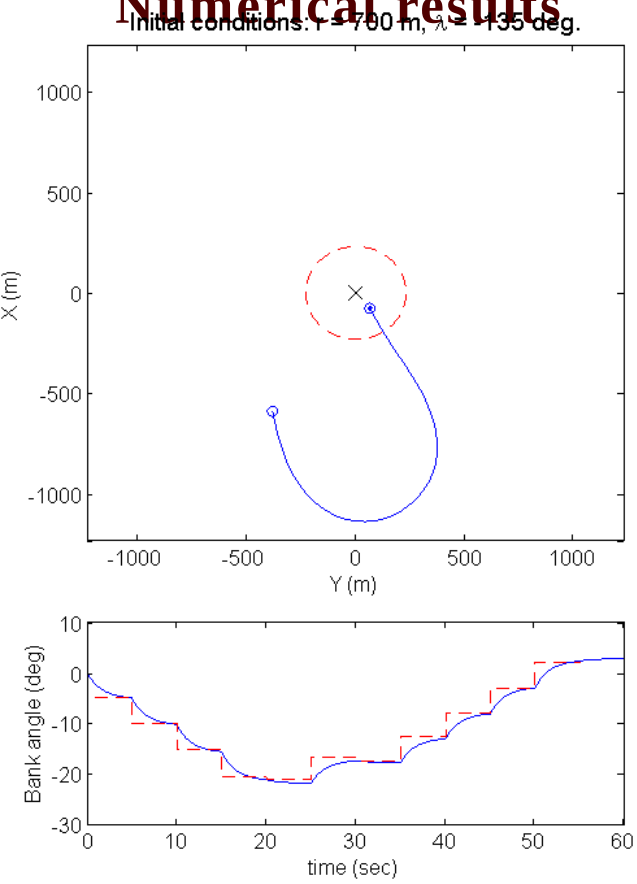


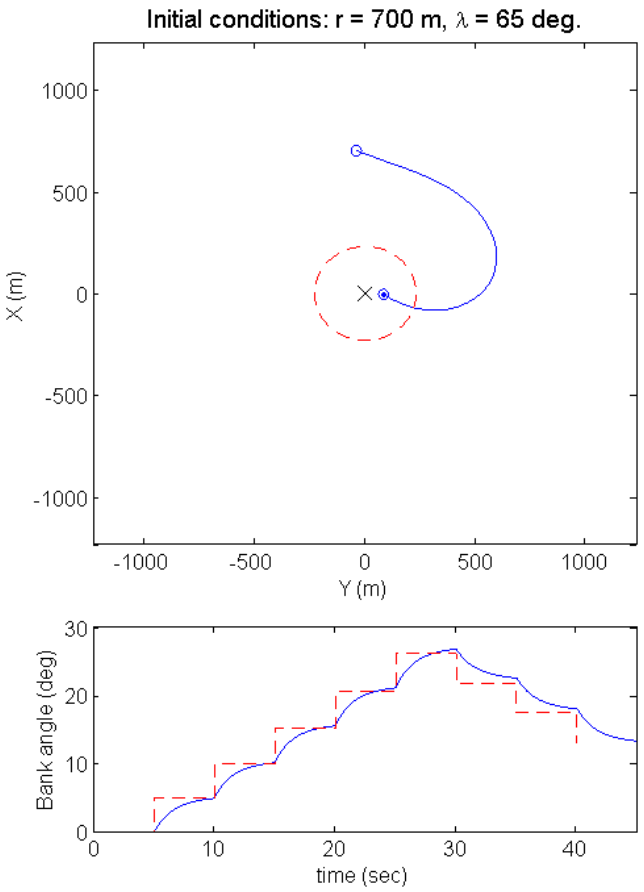


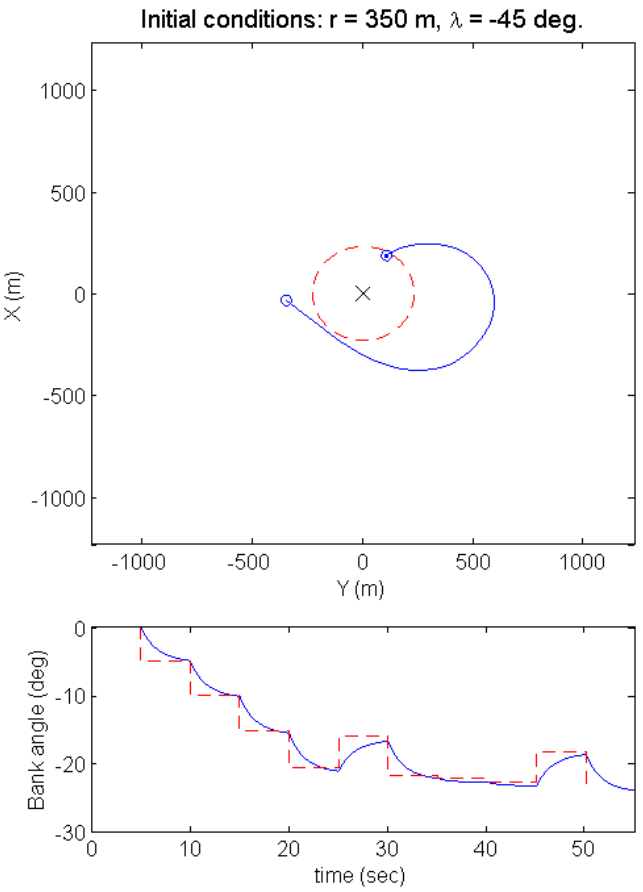


Monte Carlo results (2)

Numerical results







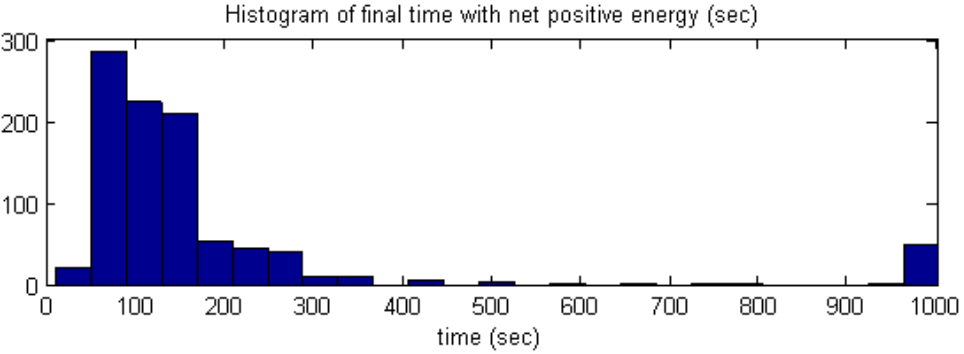
Baseline circling results (1)

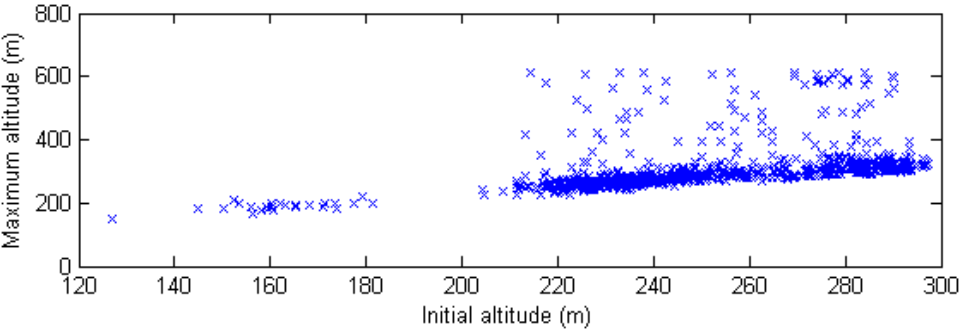
Numerical results

- Circling controller simulated for 1000 seconds
- Baseline for later work
- Plots:
  - Time before normalized energy reaches its initial value
  - Peak altitude while circling
- In 51 cases energy still increasing at t = 1000

Baseline circling results (2)

Numerical results





- Mean final time: 151.3 sec
- Median final time: 118.2 sec
- Mean change in altitude: 50.1 m
- Standard deviation: 69.9 m
- 140 altitudes show outlier behavior

Conclusions

- Novel method for guidance to thermal updraft
  - Low computational overhead in implementation
  - Somewhat model-agnostic
  - Flexibility to tailor performance via reward shaping
  - Preliminary agent succeeds in 96.2% of cases, including some initial conditions not used in learning
- Future work:
  - Disturbance tolerance
  - RL for circling flight
  - Test model-agnosticism
  - Thermal localization

Acknowledgements

- This material is based upon work supported in part by the U.S. Air Force Office of Scientific Research under contract FA9550-08-1-0038, with technical monitor Dr. Fariba Fahroo.
- It is also supported by the National Science Foundation Graduate Research Fellowship, and by Texas A&M University from the Undergraduate Summer Research Grant program.
- This support is gratefully acknowledged by the authors. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force.

References

[1] Boslough, M. B., "Autonomous dynamic soaring platform for distributed mobile sensor arrays," *Sandia National Laboratories, Sandia National Laboratories, Tech. Rep. SAND2002-1896*, 2002.

[2] *Glider Flying Handbook*, chap. Chapter 10: Soaring techniques, Flight Standards Service, Federal Aviation Administration, 2013, FAA-H-8083-13A.

[3] Wharington, J. and Herszberg, I., "Control of a high endurance unmanned air vehicle," Proceedings of the 21st ICAS Congress, 1998.

[4] Andersson, K., Kaminer, I., Dobrokhodov, V., and Cichella, V., "Thermal Centering Control for Autonomous Soaring; Stability Analysis and Flight Test Results," *Journal of Guidance, Control, and Dynamics*, Vol. 35, No. 3, 2012, pp. 963–975.

[5] Allen, M. and Lin, V., "Guidance and control of an autonomous soaring UAV," *NASATM-2007-214611*, 2007.

[6] Edwards, D. J., "Implementation details and flight test results of an autonomous soaring controller," *North Carolina State University*, 2008.

[7] Akhtar, N., "Control system development for autonomous soaring," 2010.

[8] Sheng, H., Chao, H., Coopmans, C., Han, J., McKee, M., and Chen, Y., "Low-cost UAV-based thermal infrared remote sensing: Platform, calibration and applications," *Mechatronics and Embedded Systems and Applications (MESA), 2010 IEEE/ASME International Conference on*, IEEE, 2010, pp. 38–43.

[9] Lawrance, N. R. and Sukkarieh, S., "Autonomous exploration of a wind field with a gliding aircraft," *Journal of Guidance, Control, and Dynamics*, Vol. 34, No. 3, 2011, pp. 719–733.

[10] Depenbusch, N. T. and Langelaan, J. W., "Receding horizon control for atmospheric energy harvesting by small UAVs," *AIAA Guidance, Navigation and Controls Conference, Toronto, Canada*, 2010.

[11] Allen, M. J., "Autonomous soaring for improved endurance of a small uninhabited air vehicle," Proceedings of the 43rd Aerospace Sciences Meeting, AIAA, 2005.

[12] Sutton, R. S. and Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[13] Dunn, C., Valasek, J., and Kirkpatrick, K., "Unmanned Air System Search and Localization Guidance Using Reinforcement Learning," Infotech@Aerospace 2012, 2012.

[14] "Official Arduplane Repository," https://code.google.com/p/ardupilot-mega/wiki/home?tm=6, Accessed May 24, 2013.

[15] Sim, A. G., "Flight characteristics of a modified Schweizer SGS 1-36 sailplane at low and very high angles of attacks," *NASA TP-3022*, 1990.

Backup

Dynamic model

The lateral-directional continuous-time dynamic model derived from Refs. 1 and 15 is:

$$\begin{bmatrix} \dot{\beta} \\ \dot{p} \\ \dot{r} \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} -0.1854 & 0.04069 & 0.9719 & -0.2984 \\ 9.732 & -19.49 & 2.585 & -0.2655 \\ -2.024 & 0.6734 & -0.6171 & 0.1212 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta \\ p \\ r \\ \phi \end{bmatrix} + \begin{bmatrix} 0.02569 & 0.09295 \\ -14.77 & -1.278 \\ 0.7632 & 1.799 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_a \\ \delta_r \end{bmatrix} \tag{15}$$

The longitudinal axis continuous-time dynamic model from the same sources is:

$$\begin{bmatrix} \dot{u} \\ \dot{\alpha} \\ \dot{q} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} -0.02451 & 5.938 & 0.4913 & -11.17 \\ -0.01475 & -2.187 & 0.7916 & 0.01964 \\ -0.05919 & -14.6 & 1.305 & -0.1116 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u \\ \alpha \\ q \\ \theta \end{bmatrix} + \begin{bmatrix} -0.1833 \\ -0.1432 \\ -0.9847 \\ 0 \end{bmatrix} \delta_e \tag{16}$$

The steady-state speeds and attitude, estimated from the flight test results in Ref. 15, are listed below. Unlisted values are zero in the steady-state.

- Body-axis forward speed: $U_1 = 34.4063 \ \frac{\text{m}}{\text{s}}$
- Pitch-axis Euler angle (positive nose up): $\theta_1 = -1.8873°$
- Angle of attack (positive nose up): $0.1766°$
- Body-axis vertical speed (positive down): $W_1 = 0.1061 \ \frac{\text{m}}{\text{s}}$

Q-learning background

Control policy

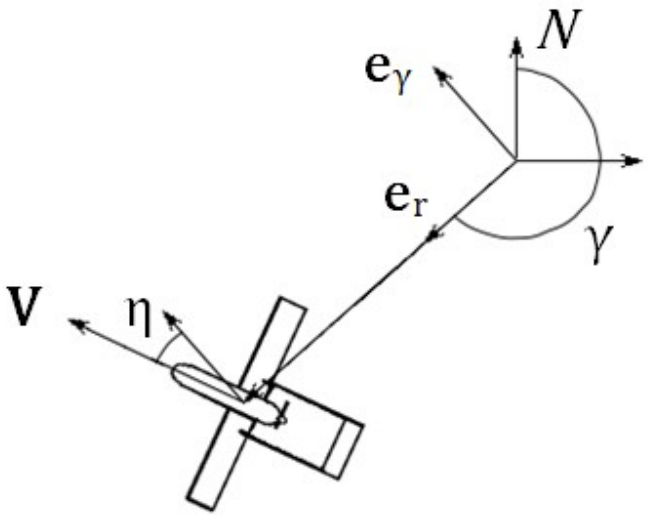
- Decision agent attempts to learn $Q^*(a_t, s_t)$ which is the true value of taking action a_t in state s_t
- Q-learning algorithm approximates Q^* by learning a mapping, designated Q
- Q guaranteed to converge to Q^* in infinite learning episodes (Watkins and Dayan, 1992)
- Convergence is policy independent
- Explicit model of system not required for learning
- ε-greedy policy used for learning
  - Decision agent takes a random (exploratory) action with frequency ε
  - ε = 1 initially and is decreased as learning progresses

Thermal circling control law

$$\mathbf{r} = r\hat{\mathbf{e}}_r$$

$$\frac{d\mathbf{r}}{dt} = \mathbf{V} = \dot{r}\hat{\mathbf{e}}_r + r\dot{\gamma}\hat{\mathbf{e}}_\gamma \qquad V = \frac{1}{2}e_\eta^2$$

$$\tan\eta = \frac{\dot{r}}{r\dot{\gamma}}$$



$$\dot{V} = e_\eta\left(\frac{r\ddot{r}\dot{\gamma} - \dot{r}^2\dot{\gamma} - r\dot{r}\ddot{\gamma}}{\dot{r}^2 + (r\dot{\gamma})^2} - \dot{\eta}_{ref}\right)$$

$$-K_L e_\eta = \frac{r\ddot{r}\dot{\gamma} - \dot{r}^2\dot{\gamma} - r\dot{r}\ddot{\gamma}}{\dot{r}^2 + (r\dot{\gamma})^2} - \dot{\eta}_{ref}$$

$$\frac{d^2\mathbf{r}}{dt^2} = (\ddot{r} - r\dot{\gamma}^2)\hat{\mathbf{e}}_r + (2\dot{r}\dot{\gamma} + r\ddot{\gamma})\hat{\mathbf{e}}_\gamma$$

$$= -g\tan\phi\hat{\mathbf{e}}_r$$

$$\ddot{r} = r\dot{\gamma}^2 - g\tan\phi$$

$$\ddot{\gamma} = \frac{-2\dot{r}\dot{\gamma}}{r}$$

$$g\tan\phi = r\dot{\gamma}^2 - \frac{U_1^2(\dot{\eta}_{ref} - K_L e_\eta) - \dot{r}^2\dot{\gamma}}{r\dot{\gamma}}$$

Updraft model (1)

Aircraft modeling and simulation

- From Allen (2005)
- w^* =2.56 m/s and zi = 660 m: scale parameters
- Z : above-ground altitude
- Updraft magnitude:

$$w_T = w^* \left(\frac{z}{z_i}\right)^{\frac{1}{3}} (1 - 1.1\frac{z}{z_i})$$

- Updraft diameter:

$$D = 0.203 \left(\frac{z}{z_i}\right)^{\frac{1}{3}} (1 - 0.25\frac{z}{z_i}) z_i$$

- Zero updraft outside the thermal assumed

Updraft model (2)

Aircraft modeling and simulation

- From Allen (2005)
- Zero updraft outside the thermal assumed
- Updraft applied to vehicle as a disturbance
- Value of flying to thermal not considered; for RL purposes, thermal size is arbitrary
- Updraft radius sized so that the SGS can circle with steady level turns of 30o at 300 m altitude
- Updraft height chosen so that energy reaches a peak within 1000 seconds of circling in most simulations

Aircraft modeling and simulation

- Bank angle PID control law:
  - $Kp = 5.5$
  - $Ki = 0.1$
  - $Kd = 7.0$

$$\delta_{pid} = -K_p(\phi - \phi_r) - K_i \int_0^t (\phi(t) - \phi_r)\mathrm{d}t - K_d\dot{\phi}$$

(subject to $-\delta_{max} < \delta_{pid} < \delta_{max}$ with $\delta max = 25o$)

- Elevator (δe) control: Linear Quadratic Regulator

$$\delta_e = \begin{bmatrix} 0.1416 & 1.925 & -0.4718 & -1.479 \end{bmatrix} \begin{bmatrix} u \\ \alpha \\ q \\ \theta \end{bmatrix}$$

Simulation framework

Aircraft modeling and simulation

- North-east-down X-Y-Z inertial coordinate system
- Standard aircraft body reference frame
- 3/2/1 Euler angle attitude parameterization through $\psi/\theta/\Phi$
- For learning Q, the state is initialized at 500 m $\leq r \leq$ 700 m, -135o $\leq \lambda \leq$ 135o, $\Phi = 0$, Z = -300 m; all other perturbed states zero.
- RL navigation agent trained on 100,000 episodes using ε-greedy policy
- Evaluated on 1,000 Monte Carlo simulations
- When the aircraft reaches the thermal, the circling controller is switched on and performance evaluated for 1,000 seconds