2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

A New Utility Function for Smooth Transition Between Exploration and Exploitation of a Wind Energy Field

Jen Jen Chung[1], Miguel Ángel Trujillo Soto[2] and Salah Sukkarieh[1]

Abstract— **This paper presents a new data driven utility function for an unmanned aerial vehicle (UAV) mapping and exploiting a wind field. The proposed utility function provides a continuous scale between exploration and exploitation which is dependent on the difference between the current platform energy level and the uncertainty along a planned path. Tests were carried out in a VICON testbed using quadrotors programmed to emulate fixed-wing aircraft. Results show a 47.7% reduction in energy gain loitering time when compared to a pure information gain approach.**

I. Introduction

Autonomous real-time energy capture for UAVs, through methods such as soaring, can provide the means to perform long endurance autonomous flight for persistent aerial systems.

Research into long endurance autonomous flight has recently gained momentum following the work into autonomous soaring carried out by [1], [2], [3]; however, the focus of this work has largely been in characterising lift-providing sources in the atmosphere [4] [5], and techniques for extracting energy from the wind [6] [7]. The additional requirement of mapping the entire wind field was considered in [8], introducing the exploration-exploitation concept for autonomous soaring. The authors used a heuristic reward function consisting of weighted energy, navigation and sampling reward components to choose the next series of roll rate and pitch rate commands. In a path following application presented in [9], the estimated altitude sink between the current UAV position and the estimated thermal centre was judged against a predefined threshold value to determine when to break from the current path and proceed to the updraft.

Currently, there lacks a continuous representation of the exploration-exploitation reward scale for autonomous soaring applications. It is hypothesised that a holistic representation of the information and energy gain will allow the computation of more efficient paths through the wind field as compared to the current discrete methods for toggling between informative paths and energy gain paths.

The primary challenge of defining such a reward function is in reconciling the energy and information gain measures. Combining the two metrics via an optimistic estimation approach was suggested in [8], where the most favourable 1σ wind change along the path segment was used to quantify the sampling reward component of the utility function. While this method is able to convert information gain into energy units, it does not provide a continuous scale between the competing objectives of exploration and exploitation, requiring additional energy and navigation reward terms to be explicitly included in the combined utility function. The authors also note that the sampling reward component is significantly smaller than the energy and navigation reward components and is effective only during tie-break situations.

The reward function presented in this paper accommodates both information gain and energy gain objectives on a continuous scale by accounting for the disparity between the current platform energy and the uncertainty of the cells to be traversed. Rather than assuming optimistic energy rewards under all circumstances, the proposed reward function is able to adjust the level of optimism based on the difference between the remaining platform energy and the map uncertainty at each sampling location along a planned path.

The proposed utility function was tested in simulation and in a quadrotor VICON testbed using a fixed-wing emulation. Results from the simulation and flight tests show a 47.7% reduction in energy gain loitering time when compared to a pure information gain utility function. A weighted combination of the two utilities produced a 19.0% reduction in loitering time. Both functions produced a lower mean estimation error and a comparable information gain rate when compared to the pure information gain case.

Section II discusses the exploration and exploitation problem for an autonomous energy capture system and presents the proposed utility function. Section III details the simple smoothing mapping technique for estimating the wind field and information gain at each point in the field. The simulation setup and results are given in Section IV, and the flight test setup and results are shown in Section V. Concluding remarks are made in Section VI.

II. Exploration and Exploitation

While the autonomous soaring problem is a relatively new and distinct manifestation of the exploration-exploitation genre, some of the fundamental principles of the classical multi-armed bandit and robot navigation scenarios can be applied when considering this problem.

The analogy of the multi-armed bandit problem is as follows: at each timestep, the player pulls one arm of a slot machine with $k \geq 2$ arms and wins a random reward drawn from an unknown probability distribution Π_j for

[1]J.J. Chung and S. Sukkarieh are with the Australian Centre for Field Robotics, Faculty of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, NSW, 2006, Australia {j.chung, salah}@acfr.usyd.edu.au

[2]M.Á. Trujillo Soto is with the Center for Advanced Aerospace Technologies (CATEC), Aerospace Technology Park of Andalusia, 41309, La Rinconada, Seville, Spain matrujillo@catec.aero

each arm j. How should the player choose which arm to pull next to maximise the total reward for a specified horizon? The problem describes the fundamental dilemma between "exploration", sampling from distributions with high uncertainty to better characterise the population parameters, and "exploitation", sampling from distributions with high expected rewards. The strategy of choosing the arm with the highest upper confidence bound was proven asymptotically optimal by [10]. This method introduced the concept of using the uncertainty measure of a distribution to contribute to the reward associated with the respective action.

The work carried out by [11] and [12] on active exploration deals directly with the problem of designing a bistable system for smooth switching between the opposing goals of "exploration" and "exploitation" in a robot navigation task using reinforcement learning. The authors introduce the variable Γ, refered to as the *attention parameter*, which is updated at each decision instance by a function of the current value of Γ and the expected change in both the exploration and exploitation states of the robot due to a particular action. The attention parameter is used to determine the level of influence of each competing objective and can thus prevent locked situations which arise when using a fixed, linear combination of competing objectives.

The utility function proposed in this paper merges and extends these concepts into the continuous, dynamic domain of a UAV mapping a wind field where the method of reinforcement learning becomes infeasible due to the curse of dimensionality. The proposed utility function for wind field mapping and exploitation is of the form,

$$reward = (\mu_W + \gamma\sigma_W), \tag{1}$$

where μ_W is the expected value of the available wind energy along a path given all the observations, σ_W is the uncertainty associated with the current sample distribution, and γ is a measure of the disparity between the current energy level and information available along the path.

$$\gamma = (1 - |E_{percent} - I_{percent}|) \times 2 - 1, \gamma \in [-1, 1]. \tag{2}$$

The variable γ plays a role similar to the *attention parameter* however its formulation and application is different to that of [12]. Γ is updated by a bistable function of the trade-off between exploration and exploitation under the current focus of attention whereas γ purely scales the difference of $E_{percent}$ and $I_{percent}$ between $[-1, 1]$ and is used as a measure of optimism to weight the variance summed in (1).

The computation of μ_W and σ_W will be dependent on the mapping estimate and uncertainty representation. Further details for the implementation in the VICON flight tests are given in Section III. The $E_{percent}$ and $I_{percent}$ values are measured by scaling the current levels of energy and information gain at a particular cell between defined upper and lower bounds. The upper and lower energy bounds are defined repectively as the initial platform energy and zero energy. In the following experiments, information gain is included in the reward for traversing a particular cell. An information gain matrix is maintained which defines the uncertainty (equivalent to the information gain) at each cell between the values of one and zero, this provides the upper and lower bounds of the information measure.

III. Wind Field Mapping

The UAV is tasked to map the wind, however the platform begins the flight with insufficient energy to explore the entire field. To complete the mission, the UAV must use the map that it is building to generate paths through the field that allow it to extract energy from the wind and continue mapping the area.

Several mapping techniques were considered and broadly fall under the categories of "parametric" and "nonparametric" methods. Parametric methods define a model of the wind and attempt to learn the parameters of the model directly from the observations. The model presented in [2] is the widely accepted standard for characterising thermal updrafts. The major limitation of parametric methods is the inability to cater for any features in the field that have not been explicitly accounted for in the model. Although not considered in this application, other features of the wind such as wind shear layers can be exploited by a UAV to increase endurance [13]. The complexity of fully specifying the equations modelling the dynamics of the wind on time and length scales appropriate to real-time soaring severely reduces the functionality of this method.

A popular nonparametric method for mapping wind fields is Gaussian Process (GP) regression. The primary benefit of using this method is that the GP model provides a continuous estimate of the mean as well as an estimate of the uncertainty at every point in the field. The reader is referred to [14] for a comprehensive study on this topic. GP regression was used in [8] and [15] to map wind fields, however in both cases it was noted that the high computational complexity of updating the GP model as new observations arrived would hinder real-time operation when the sample set became large. Sparse approximations for learning the GP online exist [16] and applying these techniques to use GP regression for real-time energy capture may be a future research direction.

For the current application, the vertical wind field is assumed to be nominally zero with localised areas of lift representing thermal updrafts. The state of the dynamic wind field at the start of each experiment is shown in Fig. 1. Initially assuming a uniformly zero field, a simple nonparametric direct mapping method was chosen to meet the real-time mapping requirements of this experiment. Smoothing was carried out over the direct observations to account for the continuous nature of the wind field and these same techniques were replicated to form an estimate of the information gain across the field.

A. Simple Smoothing

In the following experiments, a simple smoothing technique was used to map the vertical wind field. It was assumed that the UAV had equipment capable of measuring its air relative velocity and inertial position and speed, whereby

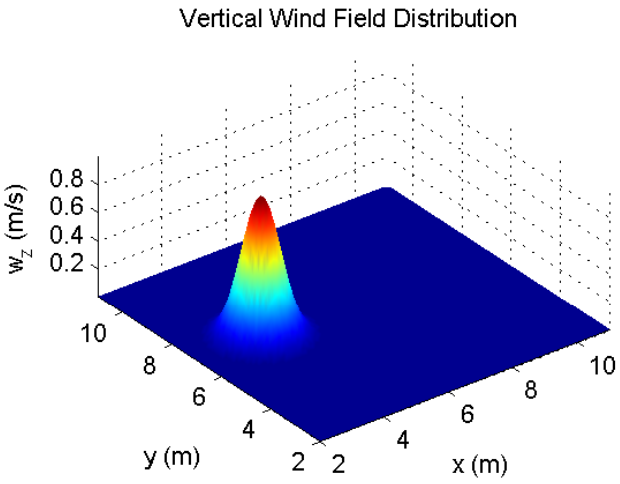


Fig. 1: Sample wind field consisting of a 2 dimensional Gaussian distribution representing a thermal vertical wind profile. In the following simulations and flight tests, the thermal centre drifts over the course of the experiment.

the difference between the air data and inertial solutions was taken as the measurement of the wind.

The wind is assumed to be continuous and smooth everywhere, thus the value of the wind field in one cell can be inferred from the value of its neighbours. The simple smoothing technique maps the direct measurements of the wind onto the grid world representation of the wind estimate and applies a Gaussian blur to the neighbours of the current cell. While this method does not perform any estimation of the wind field dynamics, it is capable of adapting the field over time as new observations arrive since the latest measurement taken at a particular cell location overwrites any previous sample taken there. It should be noted that information from prior observations is still partially retained since the smoothing technique essentially imparts some amount of information from every sample to the neighbouring cells.

The radius of the neighbourhood was chosen to cater for the shortest expected characteristic length scale of the features to be mapped in the environment. In this experiment, the mapping algorithm must be able to detect the vertical wind profile of a thermal updraft, which would show up as a hotspot amongst a relatively uniform field. Given a cell size of approximately half the expected effective updraft diameter, the radius of the neighbourhood was chosen as one cell-width such that only the direct neighbours of the current cell underwent the Gaussian blurring update at each sample time by applying a convolution matrix.

The blurring parameters A and σ^2 are, respectively, the amplitude and variance of the Gaussian distribution,

$$A\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}\right), \qquad (3)$$

which is used to define the convolution matrix. In the following experiments, $A=\sqrt{2\pi\sigma}$ and $\sigma=\text{cell-width}/2\sqrt{2}$. The values were chosen to maintain the 2σ bounds of the distribution (with some tolerance) within the radius of the neighbourhood, producing a normalised convolution matrix:

$$q=\begin{bmatrix}0.0003 & 0.0170 & 0.0003\\ 0.0170 & 0.9308 & 0.0170\\ 0.0003 & 0.0170 & 0.0003\end{bmatrix}. \qquad (4)$$

B. Sample Mean and Variance

The expected value $\mu_{W_{x,y}}$ of cell (x,y) is required to compute the reward value via the utility function described in (1). Given the set of all observations Y and a current grid world map of the wind given by matrix $\mathbf{W}$, μ_W is computed by a linear weighting of the sample mean and current cell estimate. The weights are taken from the information gain matrix $\mathbf{I}_{percent}$ as the amount of uncertainty of a partiuclar cell,

$$\mu_{W_{x,y}}=\mathbf{I}_{percent_{x,y}}\bar{Y}+\left(1-\mathbf{I}_{percent_{x,y}}\right)\mathbf{W}_{x,y}. \qquad (5)$$

The sample distribution uncertainty σ_W is taken as the estimated standard deviation of the all the observations. A modified version of the MATLAB `normfit` function was used in the following experiments to compute σ_W for the current set of obervations at each decision time.

C. Information Metric

Given the task of mapping the field, a measure of the information gain is required to evaluate the value of traversing a particular path. According to the blurring method outlined above, the value of a cell can be inferred from its neighbours due to the assumption of smoothness everywhere in the wind field. Therefore, some amount of information of the neighbouring cell values is gained at each sample time. To maintain consistency between the estimation inference and the information inference, the same smoothing technique is applied to the information gain matrix for the grid world. The information gain matrix is initialised to ones and the current sampled cell is given a value of zero while the convolution matrix in (4) is applied to its neighbouring cells at each sample time. The information gain (uncertainty) matrix and estimated wind field map from a simulation trial are given in Fig. 2.

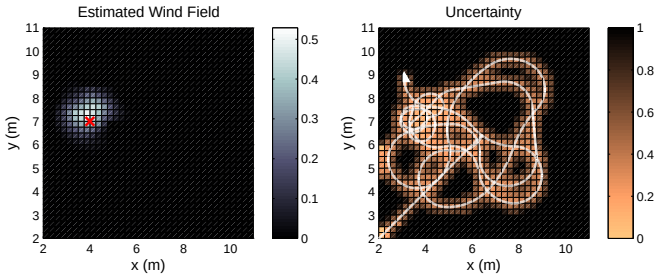


Fig. 2: Example estimated wind field map (a) and information gain matrix (b) from simulation. The actual thermal centre position is shown as a red cross in (a), the UAV path is overlaid in white and the current UAV position is shown as the white triangle in (b). The blur smooths the measured wind values and the information gained from traversing a cell across to the neighbouring cells.

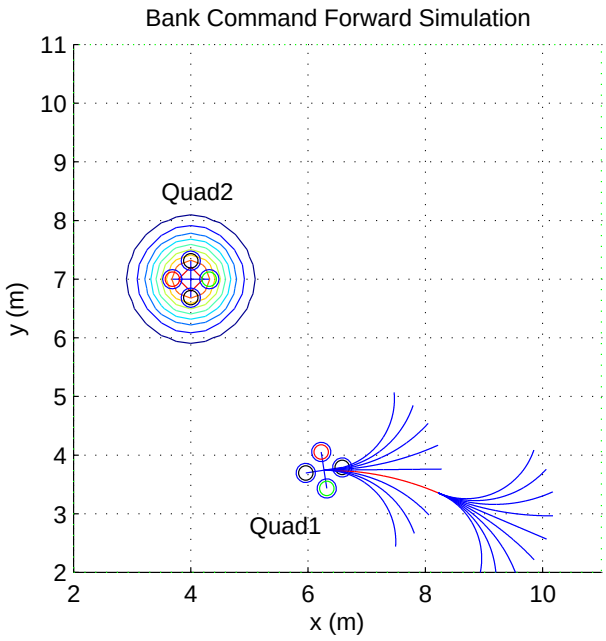


Fig. 3: A set of two planning horizons for the bank angle command set $\phi = \{-4°, -3°, -2°, -1°, 0°, 1°, 2°, 3°, 4°\}$ for Quad1. The red rotor is the left rotor and the right rotor is circled in green. At each decision time, two planning horizons are considered and commands which send the agent outside the field are heavily penalised. Only the command for the first planning horizon is executed before a new decision is made, the executed path is shown in red. Quad2 is shown with the wind energy field computed from its position which is also displayed as the underlying contour map.

IV. Experiment Setup

Experimentation was performed first in a Simulink simulation and then run in a VICON system using two quadrotors, one representing the exploring agent, the other representing an energy source. In the flight tests, data processing was performed off-platform on a desktop computer, sending the estimated wind field, updated information gain matrix and new bank commands through the VICON to the quadrotors. The energy field was defined as a vertical wind field and was computed as a Gaussian distribution centred on the second quadrotor as shown in Fig. 1. This particular distribution was chosen to model the vertical wind profile of a thermal updraft similar to the model presented in [2]. The second quadrotor was commanded to fly between two specified waypoints within the grid at a speed of 0.01m/s. The agent had no prior knowledge of the initial location of the energy source or of its subsequent motion. The amount of energy transfered between the field and the agent was computed as the value of the cell multiplied by the time spent traversing the cell. The goal of the agent was to explore and map the region of interest by exploiting energy sources found in the field to manage platform energy levels.

A. Platform Energy

The agent began the experiment with full energy (maximum altitude) and continued to lose energy at a continuous rate as it traversed the field, visualised as a constant drop in altitude. The exploring agent's altitude varied between 2m (full energy) and 1m (no energy) throughout the mission. The agent gained energy/altitude by traversing regions of the field where the contribution from the vertical wind velocity was large enough to overcome the constant energy loss. A maximum energy level was enforced so that the agent did not have sufficient energy to fully explore the field and needed to locate and loiter in areas where it expects there to be high vertical wind velocity to gain energy throughout the flight.

B. Command Modes

The experiment was designed as an emulation of a fixed-wing system capable of receiving and executing bank angle, velocity and altitude commands, similar to most commercially available fixed-wing autopilot systems. A set of nine bank angle commands, $\phi = \{-4°, -3°, -2°, -1°, 0°, 1°, 2°, 3°, 4°\}$, was available to the agent for exploring and mapping the 9m$\times$9m field. A constant velocity of 0.5m/s was commanded and this value was combined with the commanded bank angle and converted into a heading and velocity command to be sent through to the exploration quadrotor along with the altitude command computed according to the current platform energy.

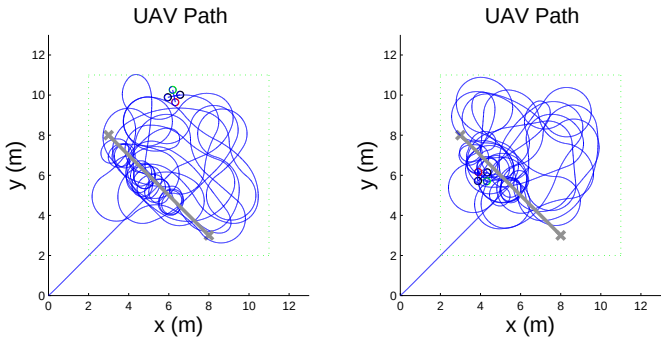
A sample set of paths derived from the bank angle set is shown in Fig. 3, each command was executed for 2s, producing an expanding tree structure similar to a state lattice. The planning horizon was two command steps, although only the first command was executed, this allowed the planner to anticipate and remove paths which would send the agent outside the field and ensure that the agent would not become "stuck" (usually near the edge of the field) with no available bank command paths after the first horizon.

The two modes of motion available to the agent were *Explore* and *Exploit*. The agent was nominally in the *Explore* mode where it executed the bank angle selected using its energy utility function. The agent switched to *Exploit* mode when its remaining energy was only sufficient in returning it to the energy target, i.e. the highest estimated vertical wind velocity location. In this mode, the agent was commanded to loiter at the energy target until its energy level exceeded a maximum threshold (80% of maximum energy) and it was returned to *Explore*. During loitering, a simple P-controller was used to adjust the commanded bank angle according to the error between the current platform heading and the bearing to the target location. This controller naturally produced an orbiting manoeuvre around the target location which was suitable to this application; for more complicated manoeuvres, a separate controller would need to be incorporated to perform energy capture.

V. Results

A. Simulation

Ten simulation trial sets were run to initially compare the performance of the proposed energy utility function to a weighted combination of the information and energy gain rewards and a pure information gain reward. For each set of three simulations, the energy source trajectory was common



(a) $\varepsilon = 0$ (b) $\varepsilon = 1$

Fig. 4: Aerial view of the exploration paths for one simulation set; the energy source drifted between (3,8) and (8,3), shown in grey, during the course of the simulation. The pure information gain approach shown in 3(a) results in more even coverage, while the new energy utility function forms a higher concentration of paths near the energy source, providing a greater number of samples near the region of high variance in the field values. The dotted border displays the edge of the field.

across the set. In (6), these three cases are given by the ε values of $\{1, 0.5, 0\}$, respectively.

$$reward = (1-\varepsilon)\mathbf{I}_{percent_{x,y}} + \varepsilon(\mu_W + \gamma\sigma_W)\Delta t. \quad (6)$$

Each trial was run for 500s, the command mode, information gain rate, information gain matrix and estimated wind field were logged to produce the following results in Table I. The mean information gain rate is the number of new cells observed per second, the percentage loitering time is measured as the amount of time spent in *Exploit* mode as a percentage of the total flight time, the percentage final uncertainty is the sum of the final information gain matrix as a percentage of the initial information gain matrix and the mean estimation error is the average m/s wind velocity error per cell of the final field estimate.

TABLE I
AGGREGATED RESULTS FROM TEN SIMULATION TRIAL SETS

	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 1$
Mean Information Rate (cells/s)	1.387	1.357	1.181
Percentage Loitering Time (%)	29.13	28.32	17.53
Percentage Final Uncertainty (%)	49.35	50.45	53.88
Mean Estimation Error (10^{-2}m/s/cell)	1.813	1.793	1.635

From Table I it can be seen that the pure information gain utility function results in the highest loitering time, and highest mean estimation error. The new energy utility function results in the lowest loitering time and lowest mean estimation error even though the mean information rate is slightly lower than the pure information gain approach. The energy utility function causes the agent to behave more conservatively than in the pure information gain approach resulting in paths that tend to remain close to the located energy source when platform energy is low (see Fig. 4). However, since the wind field appears as a hotspot field in these experiments, the greatest variance in the field values occurs near the energy source [17], thus the energy conservative approach inherently directs the agent to sample more densely in these areas, resulting in a more accurate map. The mixed utility function performed similarly to the pure information gain utility function, with slight improvements on loitering time and mean estimation error, suggesting that a larger weighting on the energy utility component may improve its performance in these categories.

B. Flight Tests

A set of three trials was carried out in the VICON testbed at Centro Avanzado de Technologías Aeroespaciales (CATEC) in Seville, Spain. The 13m×13m×4m testbed had a lateral buffering zone of 2m from each side and a flight ceiling of 3m. Data was collected at 100Hz and had a position measurement accuracy within 0.1mm. Since the quadrotor altitude was capped at 2m for the flight tests, only the 9m×9m×2m flight volume is shown in the following flight path plots.

For these experiments the agent was commanded to continue exploring if the highest estimated vertical wind velocity in the field fell below a threshold of 0.1m/s, i.e. the agent would continue to explore (presumably under "powered" flight) until it discovered a useful energy source. In all three tests the energy source, visualised by Quad2 in Fig. 5, moved between (4,7) and (4.5,9) at a constant velocity of 0.01m/s and a constant altitude of 0.5m to ensure there would be no collisions between the two quadrotors.

Figure 5 shows a period of the flight for $\varepsilon = 1$ between $t = 133$s and $t = 145$s when the exploration agent (Quad1) performs a loop around the energy source (Quad2) to gain energy before returning to explore. The complete flight paths for each trial are shown in Fig. 6. It can be seen that the energy utility function drives the agent towards the highest known energy location as its energy/altitude decreases,

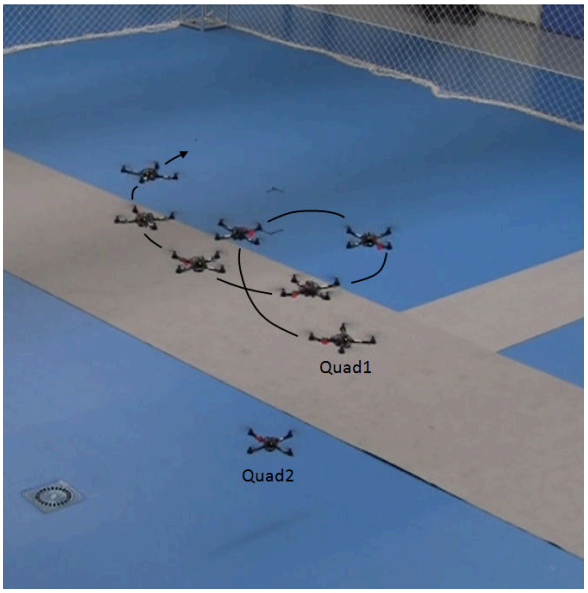
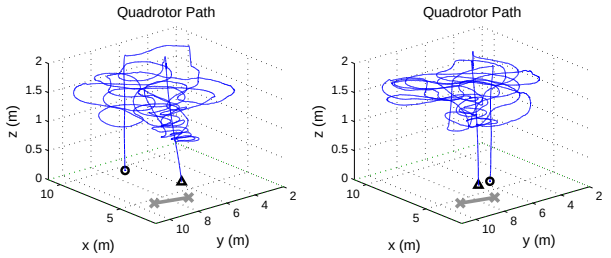
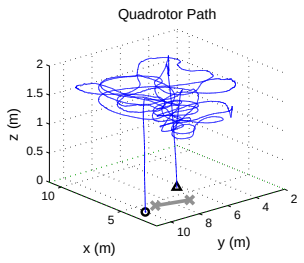


Fig. 5: A period of the flight for $\varepsilon = 1$ between $t = 133$s and $t = 145$s. The exploring agent Quad1 performs a loitering loop around the energy source Quad2 to gain energy/altitude before returning to explore the field. See the attached video of the flight tests.



(a) $\varepsilon = 0$

(b) $\varepsilon = 0.5$



(c) $\varepsilon = 1$

Fig. 6: Complete flight paths for each flight trial. The triangle and circle indicate the starting and finishing locations of Quad1, respectively. For each trial, the agent began its mission once it reached its defined starting position at (6.5,6.5,2) and the second quadrotor had reached (4,7,0.5). The path traced out by Quad2 on the ground appears as the grey line with the two commanded waypoints shown as crosses.

resulting in a relatively uniform slope towards the energy source. This behaviour is less defined in the mixed utility function case and not apparent at all in the pure information gain approach. In fact, in Fig. 6(a) there is an instance when the agent has relatively low energy but chooses to turn away from the energy source before it is forced to retrace its route to gain more energy.

The behaviour of the agent shown in Fig. 6(c) is attributed to the continuous sliding scale introduced by the γ parameter from (2) and (6). As the agent's energy decreases, γ favours visiting cells of lower uncertainty by appending a higher percentage of the sample variance to the estimated mean. Thus the reward function (6) drives the agent towards areas of high estimated mean and low uncertainty to consolidate the low platform energy.

The flight test results are given in Table II and these values agree with those obtained from simulation shown in Table I. In the flight tests, the agent performed best when directed by the mixed utility function, although the energy gain utility function produced a significantly lower loitering time compared to the other two cases. The mean information rate of flights (b), $\varepsilon = 0.5$, and (c), $\varepsilon = 1$, are 22.6% and 8.41% higher than in case (a), $\varepsilon = 0$. The mean estimation errors are also (b) 21.1% and (c) 4.72% lower than for case (a) showing that not only is the agent sampling in new areas, it is also directed to sample in the higher variance areas around the energy source to produce a more accurate map of the field.

The energy and energy mode values are plotted in Fig. 7. The energy mode value indicates whether the agent is in

TABLE II
FLIGHT TRIAL RESULTS

	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 1$
Mean Information Rate (cells/s)	1.316	1.614	1.427
Percentage Loitering Time (%)	24.89	20.17	13.03
Percentage Final Uncertainty (%)	65.09	56.67	61.42
Mean Estimation Error (10^{-2}m/s/cell)	1.823	1.439	1.737

Explore mode (mode $= 0$) or *Exploit* mode (mode $= 1$). Despite having discovered the location of the energy source much later than in the first two cases (and requiring a period of "powered flight" between $t = 124$s and $t = 147$s), the percentage loitering time is highest in the pure information gain case at 24.89%. The percentage loitering time achieved by the new energy utility function is 13.03%, and the mixed utility function achieved a 20.17% loitering time. The effect of the energy utility function is visible through the sequential rises in energy while the agent is in *Explore* mode, indicating that the chosen exploration paths periodically traverse through the energy source to compensate for the constant energy loss. As the agent's energy drops, the frequency of the energy peaks rises due to the change in the agent's optimism. In the linear weighted utility case, the primary influence of the energy utility component was to maintain proximity to the energy source when low platform energy was detected so that it was able to reach the area of lifting air faster during *Exploit* mode. To compare, the average time to reach the energy source once *Exploit* mode was triggered was (a) 9.68s, (b) 3.62s and (c) 6.07s, therefore the savings in loitering time between cases (a) and (b) arise mainly from the initial time taken to reach and relocalise the energy source.

VI. CONCLUSIONS

The results presented in this paper show that the continuous scale between exploration and exploitation achieved by the proposed energy utility function is capable of directing a simultaneously exploring and energy-capturing agent to maintain an appropriate proximity to the energy source as the available platform energy changes whilst still meeting the exploration objective of mapping the field. The information gathering performance is comparable to that of a pure information gain utility function while map estimation is improved with the new energy utility function. The proposed utility function is particularly suited to mapping hotspot fields where the energy sources are characterised by small concentrations of high variance values.

An area of future work is to extend the approach presented in this paper to problems where there is stochasticity in both the vehicle dynamics and the wind field by reformulating the γ variable to incorporate the relevant uncertainty values.

REFERENCES

[1] J. Wharington, "Autonomous control of a soaring aircraft by reinforcement learning," Ph.D. dissertation, Department of Aerospace Engineering, Royal Melbourne Institute of Technology, 1998.

[2] M. J. Allen, "Autonomous soaring for improved endurance of small uninhabited air vehicle," in *43rd AIAA Aerospace Sciences Meeting and Exhibit*, 2005.

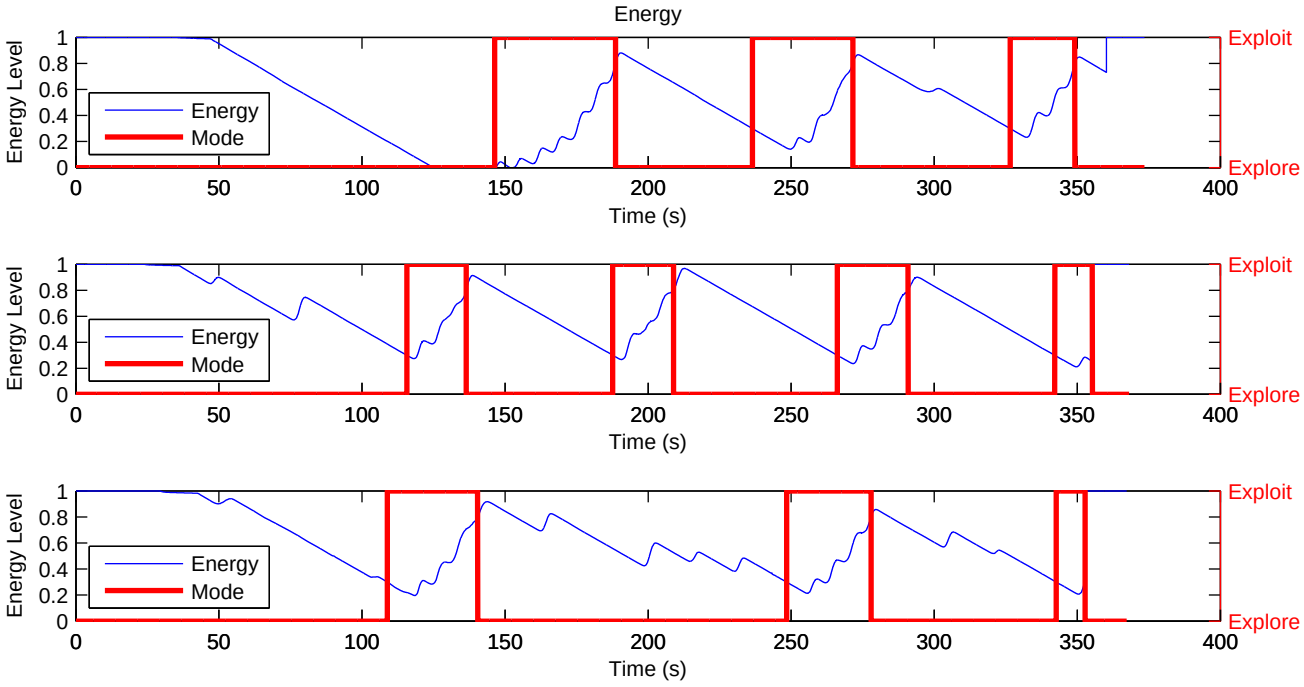


Fig. 7: Energy and energy mode values for the three flight tests, $\varepsilon = 0$, $\varepsilon = 0.5$ and $\varepsilon = 1$ in descending order. The steps represent the change from *Explore*, mode $= 0$, to *Exploit*, mode $= 1$.

[3] J. W. Langelaan, "Long distance/duration trajectory optimization for small UAVs," in *2007 AIAA Guidance, Navigation and Control Conference and Exhibit*, 2007.

[4] M. K. Hook, D. A. Findlay, A. G. Purcell, and R. T. Watkin, "Autonomous soaring," in *2007 Institution of Engineering and Technology Conference on Autonomous Systems*. IET, 2007, pp. 1–6.

[5] M. J. Allen, "Updraft model for development of autonomous soaring uninhabited air vehicles," in *44th AIAA Aerospace Sciences Meeting and Exhibit*, 2006.

[6] G. C. Bower, T. C. Flanzer, A. D. Naiman, and S. Saripalli, "Dynamic environment mapping for autonomous thermal soaring," in *AIAA Guidance, Navigation and Control Conference*, 2010.

[7] O. K. Ariff and T. H. Go, "Waypoint navigation of small-scale UAV incorporating dynamic soaring," *Journal of Navigation*, vol. 64, no. 1, pp. 29–44, 2011.

[8] N. R. J. Lawrance and S. Sukkarieh, "Autonomous exploration of a wind field with a gliding aircraft," *Journal of Guidance, Control, and Dynamics*, vol. 34, no. 3, 2011.

[9] C. Antal, O. Granichin, and S. Levi, "Adaptive autonomous soaring of multiple UAVs using simultaneous perturbation stochastic approximation," in *49th IEEE Conference on Decision and Control*, 2010.

[10] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *The Annals of Statistics*, vol. 15, no. 3, pp. 1091–1114, 1987.

[11] S. Thrun and K. Möller, "Active exploration in dynamic environments," in *NIPS*, 1991, pp. 531–538.

[12] S. B. Thrun, "Efficient exploration in reinforcement learning," School of Computer Science, Carnegie-Mellon University, Tech. Rep. CMU-CS-92-102, January 1992.

[13] Y. J. Zhao, "Optimal patterns of glider dynamic soaring," *Optimal control applications and methods*, vol. 25, no. 2, pp. 67–89, 2004.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[15] L. Csató, D. Cornford, and M. Opper, "Online learning of wind-field models," in *International Conference on Artificial Neural Networks*, 2001, pp. 300–307.

[16] L. Csató and M. Opper, "Sparse on-line gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.

[17] K. H. Low, J. M. Dolan, and P. Khosla, "Information-theoretic multi-robot adaptive exploration and mapping of environmental hotspot fields," in *ESSA 2009: Workshop on Sensor Networks for Earth and Space Sciences Applications*, 2009.