UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning to soar using atmospheric thermals

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics

by

Gautam Reddy Nallamala

Committee in charge:

Professor Massimo Vergassola, Chair
Professor Garrison W. Cottrell
Professor Patrick H. Diamond
Professor Thomas Murphy
Professor Terrence J. Sejnowski

2019

The dissertation of Gautam Reddy Nallamala is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'm thankful to my fellow graduate students for their encouragement and support. I thank my co-authors Jerome Wong-Ng, Terry Sejnowski and Antonio Celani for their ideas and advice. My warmest gratitude to my advisor, Massimo Vergassola, for his constant support throughout the journey, for giving me the confidence and freedom to pursue my own little naive ideas, and for getting me back on track when I strayed too far from what was needed to be done. I thank my wonderful parents, who encouraged me to pursue my dreams and who gladly gave me support when it was most needed. Finally, I thank my family of friends in San Diego who made my Ph.D years enjoyable and truly memorable. My warmest gratitude to Pek Ieong, a source of encouragement and inspiration.

Chapter 1, in full, is a reprint of the material as it appears in Reddy G., Celani A., Sejnowski T. J. & Vergassola M., Learning to soar in turbulent environments, *Proc. Natl. Acad. Sci.*, 113-33:4877-4884, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Reddy G., Wong Ng J., Celani A., Sejnowski T. J. & Vergassola M., Glider soaring via reinforcement learning in the field, *Nature*, Vol. 562, pp. 236-239, 2018. The dissertation author was the primary investigator and author of this paper.

VITA

2009-2013	B.Tech. in Engineering Physics *with honors*, Indian Institute of Technology, Bombay, India.
2013-	Ph. D. in Physics, Advisor: Prof. Massimo Vergassola, University of California San Diego.

PUBLICATIONS

Carballo-Pacheco M., Desponds J., Gavrilchenko T., Mayer A., Prizak R., Reddy G., Nemenman I., & Mora T., Receptor crosstalk improves concentration sensing of multiple ligands, *Phys. Rev. E*, Vol. 99, 022423, 2019.

Ryali C., Reddy G. & Yu A., Demystifying excessively volatile human learning: a Bayesian persistent prior and a neural approximation, *Advances in Neural Information Processing Systems (NIPS)*, 31, 2018.

Reddy G., Wong Ng J., Celani A., Sejnowski T. J. & Vergassola M., Glider soaring via reinforcement learning in the field, *Nature*, Vol. 562, pp. 236-239, 2018.

Reddy G., Zak J., Vergassola M. & Murthy V., Antagonism in olfactory receptor neurons and its implications for the perception of odor mixtures, *eLife*, 7:e34958, 2018.

Reddy G., Celani A., Sejnowski T. J. & Vergassola M., Learning to soar in turbulent environments, *Proc. Natl. Acad. Sci.*, 113-33:4877-4884, 2016.

Reddy G., Celani A. & Vergassola M., Infomax strategies for an optimal balance between exploration and exploitation, *J. Stat. Phys.*, 163:1454-1476, 2016.

ABSTRACT OF THE DISSERTATION

Learning to soar using atmospheric thermals

by

Gautam Reddy Nallamala

Doctor of Philosophy in Physics

University of California San Diego, 2019

Professor Massimo Vergassola, Chair

Soaring birds often rely on ascending thermal plumes (thermals) in the atmosphere as they search for prey or migrate across large distances. How soaring birds find and navigate thermals is unknown. This is a scenario where experiments are difficult to control and the strategies used by birds are difficult to infer. In this work, I used modern methods from artificial intelligence as tools to generate hypotheses for the strategies and mechanosensory cues that birds may use in order to soar effectively.

In Chapter 1, I describe how a technique from artificial intelligence, namely, reinforcement learning, is used to train virtual gliders with bird-like aerodynamic properties to navigate simulated convective turbulent flows. By experimenting with the learning environment, we find that gliders

need to sense and respond to two cues in order to soaring effectively: the vertical wind acceleration and the velocity differences across the wings. The learning process also yields a strategy for soaring within thermals that relies on these two cues.

In Chapter 2, I describe the details of how lessons from the simulations were used to teach gliders to navigate atmospheric thermals in the field. Gliders of two-meter wingspan were equipped with a flight controller that precisely controlled the bank angle and pitch, modulating these at intervals with the aim of gaining as much lift as possible. A navigational strategy was determined solely from the gliders pooled experiences collected over several days in the field. The strategy relies on methods to accurately estimate the local vertical wind accelerations and the roll-wise torques on the glider, which serve as navigational cues. I show that vertical wind accelerations and roll-wise torques are effective mechanosensory cues for soaring and provide a navigational strategy applicable to autonomous soaring vehicles.

Chapter 1

Learning to soar in simulated turbulent environments

Migrating birds and gliders use upward wind currents in the atmosphere to gain height while minimizing the energy cost of propulsion by the flapping of the wings or engines [1, 2]. This mode of flight, called soaring, has been observed in a variety of birds. For instance, birds of prey use soaring to maintain an elevated vantage point in their search for food [3] ; migrating storks exploit soaring to cover large distances in their quest for greener pastures [4]. Different forms of soaring have been observed. Of particular interest here is thermal soaring, where a bird gains height by using warm air currents (thermals) formed in the atmospheric boundary layer. For both birds and gliders, a crucial part of thermal soaring is to identify a thermal and to find and maintain its core, where the lift is typically the largest. Once migratory birds have climbed up to the top of a thermal, they glide down to the next thermal and repeat the process, a migration strategy that strongly reduces energy costs [4]. Soaring strategies are also important for technological applications, namely the development of autonomous gliders that can fly large distances with minimal energy consumption [5].

Thermals arise as ascending convective plumes driven by the temperature gradient created

due to the heating of the earth's surface by the sun [6]. Hydrodynamic instabilities and processes that lead to the formation of a thermal inevitably give rise to a turbulent environment characterized by strong, erratic fluctuations [7, 8]. Birds or gliders attempting to find and maintain a thermal face the challenge of identifying the potentially long-lived and large-scale wind fluctuations amidst a noisy turbulent background. The structure of turbulence is highly complex, with fluctuations occurring at many different scales and long-ranged correlations in space and time [9, 10]. We thereby expect non-trivial correlations between the large-scale convective plumes and the locally fluctuating quantities. Thermal soaring is a particularly interesting example of navigation within turbulent flows, since the velocity amplitudes of a glider or bird are of the same order of magnitude as the fluctuating flow they are immersed in.

It has been frequently observed and attested by glider pilots that birds are able to identify and navigate thermals more accurately than human pilots endowed with modern instrumentation [11]. It is an open problem, though, what sensorimotor cues are available to birds and how they are exploited, which constitutes a major motivation for the present study.

An active agent navigating a turbulent environment has to gather information about the fluctuating flow while simultaneously using the flow to ascend. Thus, the problem faced by the agent bears similarities to the general problem of balancing exploration and exploitation in uncertain environments, which has been well-studied in the reinforcement learning framework [12]. The general idea of reinforcement learning is to selectively reinforce actions that are highly rewarding and thereby have the reinforced actions chosen when the situation reoccurs. The solution to a reinforcement learning problem typically yields a behavioral policy that is approximately optimal, where optimality is defined in the sense of maximizing the reward function used to train the agent.

The previous description suggests that reinforcement learning methods are poised to deliver effective strategies of soaring flight. Past applications are indeed promising yet they have considered the soaring problem in unrealistically simplified situations, with no turbulence or

with fluctuations modeled as Gaussian white noise. Ref. [13] considered the learning problem associated with finding the center of a stationary thermal without turbulence, and used a neural-based algorithm to recover the empirical rules proposed by Reichmann [14] to locate the core of the thermal. Other attempts [15, 16] have used neural networks and Q-learning to find strategies to center a turbulence-free thermal. Akos et al. [17] show that these simple rules fail even in the presence of modest velocity fluctuations modeled as Gaussian white noise, and express the need for strategies that could work in realistic turbulent flows.

Here, we enforce realistic aerodynamic constraints on the flight of gliders, and train them in complex turbulent environments by using reinforcement learning algorithms. We show that the glider finds an effective strategy for soaring and we identify sensorimotor cues that are most relevant for guiding turbulent navigation. Our soaring strategy is effective even in the presence of strong fluctuations. The predicted strategy of flight lends itself to field experiments with remote-controlled gliders and to comparisons with the behavior of soaring birds.

1.1 Models

We first describe the models used for the simulation of the atmospheric boundary layer flow, the mechanics of flight and the reinforcement learning algorithms that we have employed. The next Section will then present the corresponding results.

1.1.1 Modeling the turbulent environment

Conditions ideal for thermal soaring typically occur during a sunny day, when a strong temperature gradient between the surface of the Earth and the top of the atmospheric boundary layer creates convective thermals [7, 8]. The soaring of birds and gliders primarily occurs within this convective boundary layer. The mechanical and thermal forces within the boundary layer generate turbulence characterized by strongly fluctuating wind velocities.

Key physical aspects of the flow in the convective boundary layer are governed by Rayleigh-Bénard convection (see [9] for a review). The corresponding equations are derived from the Navier-Stokes equations with coupled temperature and velocity fields simplified using the Boussinesq approximation. The dimensionless Rayleigh-Bénard equations read

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} = -\nabla P + \left(\frac{\mathrm{Pr}}{\mathrm{Ra}}\right)^{1/2}\nabla^2\boldsymbol{u} + \theta\hat{\boldsymbol{z}}\,, \tag{1.1}$$

$$\frac{\partial \theta}{\partial t} + \boldsymbol{u}\cdot\nabla\theta = \frac{1}{(\mathrm{Pr}\,\mathrm{Ra})^{1/2}}\nabla^2\theta \tag{1.2}$$

where $\boldsymbol{u}, \theta$ and P are the velocity, temperature and pressure fields, respectively. The vertical direction coincides with the z-axis. The temperature appears in the dynamics of the velocity field as a buoyant forcing term. The equations contain two dimensionless quantities that determine the qualitative behavior of the flow : The Rayleigh number, Ra, and the Prandtl number, Pr. When Ra is beyond a critical value $\sim 10^3$, the thermally generated buoyancy drives the flow towards instability. In this regime, the flow is characterized by large-scale convective cells and turbulent eddies at every length scale. In the atmosphere, the Rayleigh number can reach up to Ra = $10^{15} - 10^{20}$. In such high-Rayleigh-number regimes, the flow is strongly turbulent and numerical simulations of convection in the atmosphere are thus plagued by the same limitations of simulating fully developed turbulent flows. We performed direct numerical simulations of Rayleigh-Bénard convection at Ra = 10^8 using the Gerris Flow Solver [18] (see Appendix A for more details about the grid and the numerical scheme). Our test arena is a three-dimensional cubical box of side length 1 km in physical units. We impose periodic boundary conditions on the lateral walls and no-slip on the floor and the ceiling of the box. The floor is fixed at a high temperature (which is rescaled to $\theta = 1$) and the ceiling is fixed at $\theta = 0$.

A small, random perturbation in the flow quickly leads to an instability and to the formation of coherent thermal plumes within the chamber. Snapshots of the velocity and temperature fields at the statistically stationary state are shown in Fig. 1.1A. The statistical properties of the flow

are consistent with those observed in previous works [19, 20], particularly the Nusselt number (which measures the ratio of convective to conductive heat transfer) and the mean temperature and velocity field profiles (see Figure A.1) in the Appendix A).

To test the robustness of our learned policies of flight with respect to the modeling of turbulence, we also considered an alternative to the Rayleigh-Bénard flow. Specifically, we considered a kinematic model of turbulence that extends the one in [21] to the inhomogeneous case relevant for the atmospheric boundary layer (see Appendix A). Results for the kinematic model confirm the robustness of our conclusions and the learned policy has similar features in both flows (Figure A.3). Below, we shall focus on the simulations of the Rayleigh-Bénard flow described above.

1.1.2 Glider mechanics

A bird or glider flying in the flow described above with a fixed, stretched-out wing is safely assumed to be in mechanical equilibrium, except for centripetal forces while turning [22, 23]. A glider with weight W traveling with velocity $\boldsymbol{v}$ experiences a lift force L perpendicular to its velocity and a drag force D antiparallel to its velocity (see Fig. 1.1C for a force body diagram). The glider has no engine and thus generates no thrust. The magnitudes of the lift and the drag depend on the speed v, the angle of attack α, the density of air ρ and the surface area S of the wing as : $L = \frac{1}{2}\rho S v^2 C_L(\alpha)$ and $D = \frac{1}{2}\rho S v^2 C_D(\alpha)$. The glide angle γ, which is the angle between the velocity and its projection on the horizontal, determines the ratio of the climb rate $v_c(<0)$ and the horizontal speed $v_\perp$. Balancing the forces on the glider, and accounting for the centripetal acceleration, the velocity of the glider and its turning rate are obtained :

$$\tan\gamma = \frac{-v_c}{v_\perp} = \frac{D}{L\cos\mu} = \frac{C_D(\alpha)}{C_L(\alpha)\cos\mu}; \tag{1.3}$$

$$\dot{y} = g\cos\gamma\tan\mu; \quad v^2 = \frac{2mg\sin\gamma}{\rho S C_D(\alpha)}. \tag{1.4}$$

Here, $\ddot{y}$ is the centripetal acceleration. The ratio mg/S is called the wing loading of the glider [22]. The kinematics of a glider is therefore set by the wing loading and the dependence of the lift and the drag coefficients on the angle of attack. The general features of the lift and drag coefficient curves for a typical symmetric airfoil are described in [24]; the resulting dependence of the velocity on the angle of attack is shown in Fig. 1.1B. The glider can be maneuvered by controlling the angle of attack, which changes the speed and climb rate of the glider, or by banking the glider to turn.

1.1.3 The learning algorithm

To identify effective strategies of soaring flight in turbulent flows, we used the reinforcement learning algorithm SARSA [12]. Historically, the algorithm was inspired by the theory of animal learning, and its model-free nature allows for learning previously unknown strategies driven by feedback on performance [25].

Reinforcement learning problems are typically posed in the framework of a Markov Decision Process (MDP). In an MDP, the agent traverses a state space with transition probabilities that depend only on the current state s and the immediate next state s', as for a Markov process. The transition probabilities can be influenced by taking actions at each time step. After every action, the agent is given some reward $r(s,s',a)$, which depends on the states s and s' and the chosen action a. The ultimate goal of reinforcement learning algorithms is to find the optimal policy π^*, i.e. find the probability of choosing action a given the state s. The optimal policy maximizes for each state s the sum of discounted future rewards $V_{\pi_s^a}(s) = \langle r_0 \rangle + \beta \langle r_1 \rangle + \beta^2 \langle r_2 \rangle + \ldots$, where $\langle r_i \rangle$ is the expected reward after i steps, β is the discount factor ($0 < \beta < 1$) and the sum above obviously depends on the policy π_s^a. When β is close to zero, the optimal policy greedily maximizes the expected immediate reward, leading to a purely exploitative strategy. As β gets closer to unity, later rewards contribute significantly and more exploratory strategies are preferred.

The SARSA algorithm finds the optimal policy by estimating for every state-action pair

its Q-function defined as the expected sum of future rewards given the current state s and the action a. At each step, the Q-function is updated as

$$Q(s,a) \to Q(s,a) + \eta(r + \beta Q(s',a') - Q(s,a)), \tag{1.5}$$

where r is the received reward and η is the learning rate. The update is made online and does not require any prior model of the flow or the flight. This feature is particularly relevant in modeling decision-making processes in animals. In the brain, reinforcement learning depends on a related reward prediction error, which is represented by a system of neurons that use dopamine as their neurotransmitter [26]. When the algorithm is close to convergence, the Q-function approaches the solution to Bellman's dynamic programming equations [12]. The policy π_s^a, which encodes the probability of choosing action a at state s, approaches the optimal one π^* and is obtained from the Q-function via a Boltzmann-like expression :

$$\pi_s^a \propto \exp\big(-\hat{Q}(s,a)/\tau_{\text{temp}}\big), \tag{1.6}$$

$$\hat{Q}(s,a) = \frac{\max_{a'} Q(s,a') - Q(s,a)}{\max_{a'} Q(s,a') - \min_{a'} Q(s,a')}. \tag{1.7}$$

Here, τ_{temp} is an effective"temperature" : when $\tau_{\text{temp}} \gg 1$, actions are only weakly dependent on the associated Q-function ; conversely, for τ_{temp} small, the policy greedily chooses the action with the largest Q. The temperature parameter is initially chosen large and lowered as training progresses to create an annealing effect, thereby preventing the policy from getting stuck in local extrema. Parameters used in our simulations can be found in Table A.1.

In the sequel, we shall qualify the policy identified by SARSA as optimal. It should be understood though that the SARSA algorithm (as other reinforcement learning algorithms) typically identifies an approximately optimal policy and "approximately" is skipped only for the sake of conciseness.

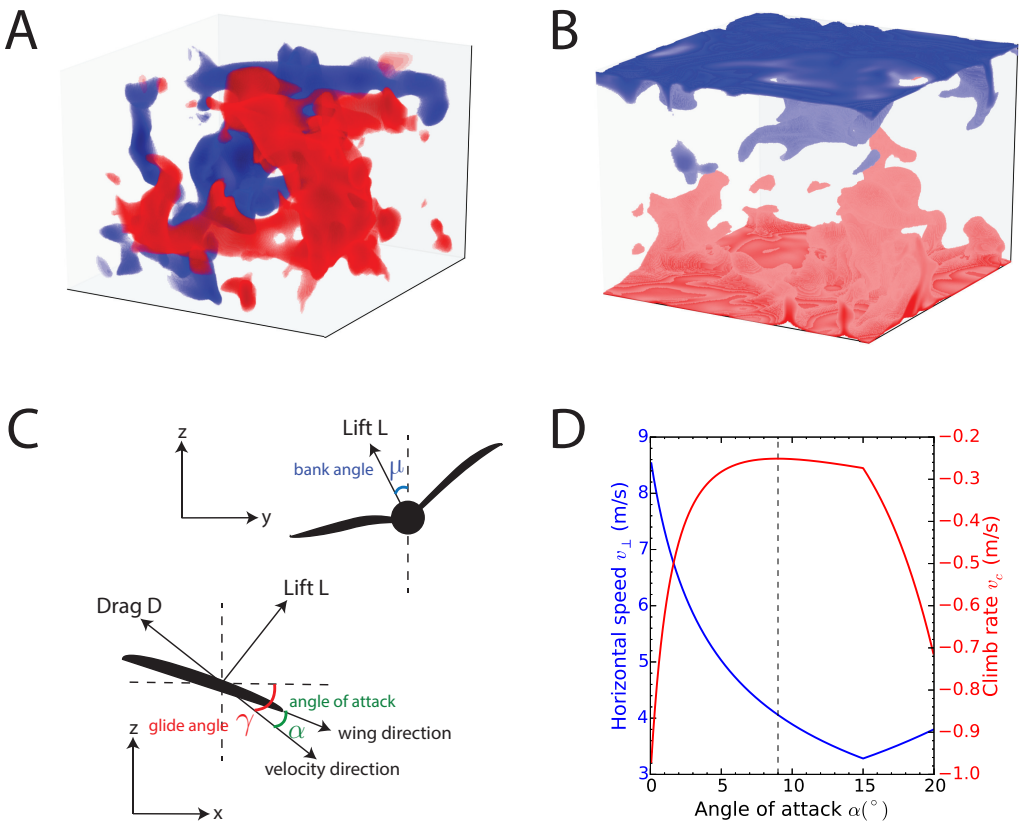

Figure 1.1: Snapshots of the vertical velocity (A) and the temperature fields (B) in our simulations of Rayleigh-Bénard convection. (C) : The force-body diagram of flight with no thrust (D): The range of horizontal speeds and climb rates accessible by controlling the angle of attack.

1.2 Results

1.2.1 Sensorimotor cues and reward function for effective learning

Key aspects of the learning for the soaring problem are the sensorimotor cues that the glider can sense (state space), and the choice of the reward used to train the glider to ascend quickly. As the state and action spaces are continuous and high-dimensional, it is necessary to discretize them, which we realize here by a standard tile coding scheme [12]. The height ascended per trial, averaged over different realizations of the flow, serves as our performance criterion.

The glider is allowed control over its angle of attack and its bank angle (see Fig. 1.1B). Control over the angle of attack features two regimes : (1) at small angles of attack, the horizontal

speed is large and the climb rate is small (the glider sinks quickly) ; (2) at large angles of attack but below the stall angle, the horizontal speed is small whereas the climb rate is large. The bank angle controls the heading of the glider, and we allow for a range of variation between -15° and 15°. Exploring various possibilities, we found that three actions are minimally sufficient : increasing, decreasing or preserving the angle of attack and the bank angle. The angle of attack and bank angle were incremented/decremented in steps of 2.5° and 5° respectively. In summary, the glider can choose 3^2 possible actions to control its navigation in response to the sensorimotor cues described hereafter.

Our rationale in the choice of the state space was trying to minimize biological or electronic sensory devices necessary for control. We tested different combinations of local sensorimotor cues that could be indicative of the existence of a thermal. These were the vertical wind velocity $\boldsymbol{u}_z$, the vertical wind acceleration $\boldsymbol{a}_z$, torques τ, the local temperature θ, and their sixteen possible combinations. Namely, if $\boldsymbol{u}$ denotes the local windspeed, we define the wind acceleration as $\boldsymbol{a}_z = (\boldsymbol{u}_z^{(t)} - \boldsymbol{u}_z^{(t-1)})/\Delta t$ and the "torques" as $\tau = (\boldsymbol{u}_{z+} - \boldsymbol{u}_{z-})l$, where $\boldsymbol{u}_{z+}$ and $\boldsymbol{u}_{z-}$ are the vertical wind velocities at the left and the right wing, l is the wingspan of the glider and Δt is the step used for time discretization (see below). After experimentation with various architectures, we found that a look-up table structure with three states per observable, corresponding to positive high, negative high and small values, ensures good performance. The chosen thresholds, $\boldsymbol{a}_z^{\text{thresh}}$ and τ^{thresh}, that demarcate the large and small values in our tile coding scheme are listed in Table A.1.

As for the reward function, we found that a purely global reward, i.e. awarded at the end of a trial without any local guidance, does not propagate easily to early state-action pairs for realistically long trials. Eligibility traces [12], which maintain a memory of past state-action pairs and their rewards, did not alleviate the issue. For gliders or migrating birds, a fall can be extremely disadvantageous and we account for this by having a glider which touches the surface receive a large negative reward as a penalty. After a broad exploration of various choices, we

heuristically found that best soaring performances are obtained by a local-in-time reward that linearly combines the vertical wind velocity and the wind acceleration achieved at the subsequent time step.

1.2.2 Flight training

The glider is first trained on a set of trials and its performance is then tested on 500 trials. Trials consist of independent statistical realizations of the turbulent flow. The glider flight is discretized by time steps $\Delta t = 1s$, which is an estimate for the control times of the glider and the time-scales of the turbulent eddies at the size of the glider. Each trial lasts for two and a half minutes, which is roughly half the relaxation time of the large-scale convective flow at steady state. The duration captures the order of magnitude of the typical time, ~ 10 mins, for birds to reach the base of the clouds.

The velocity relative to the ground of the glider is $\boldsymbol{u} + \boldsymbol{v}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the contributions due to the wind and the glider velocity, respectively. If u_{rms} is the root-mean-squared speed of the flow and v_{glider} is the typical airspeed of the glider, we introduce their dimensionless ratio $\hat{u}_{\text{rms}} = u_{\text{rms}}/v_{\text{glider}}$. At small $\hat{u}_{\text{rms}}$, fluctuations are weak. Conversely, at large $\hat{u}_{\text{rms}}$, the glider has less time to react to rapidly changing velocities, i.e. the environment is strongly fluctuating. Moreover, in that regime the glider is carried away by the flow and the amount of control the glider has over its trajectory is reduced. We expect that the policy of flight learned by the glider will differ between the regimes of weak and strong fluctuations.

1.2.3 Learning in different flow regimes

A qualitative sense of the efficiency of the training in a fluctuating regime is illustrated in Fig. 1.2. The trajectories go from random paths to the spirals that are characteristic of the thermal soaring flights of birds and gliders. Fig. 1.3A quantifies the significant improvement

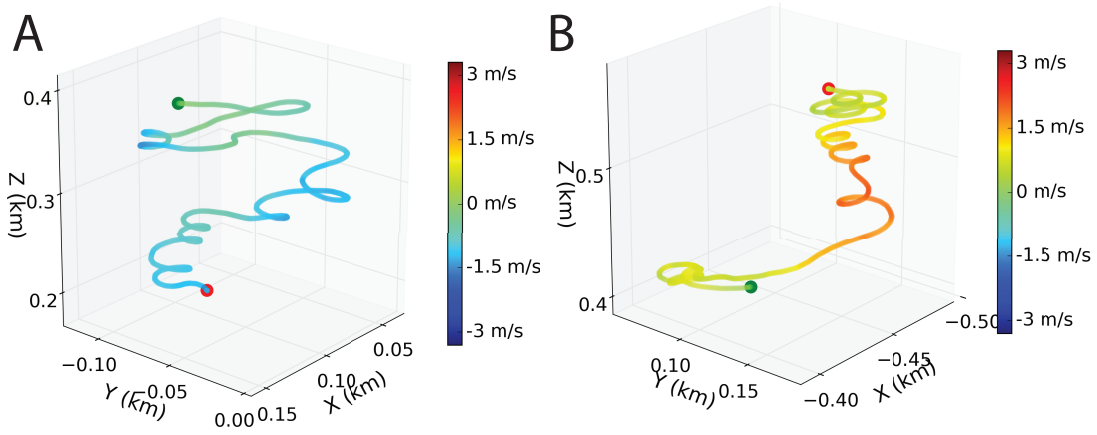

Figure 1.2: Typical trajectories of an untrained (A) and a trained (B) glider flying within a Rayleigh-Bénard turbulent flow, as shown in Fig. 1.1. The green and red dots indicate the start and the end points of the trajectory, respectively.

in performance due to training and shows that training for a few hundred trials suffices for convergence with negligible overfitting for larger training sets. To compare performance in flows of different mean speeds, we train and test gliders in flows with varying u_{rms}. Fig. 1.3B shows the gain in height as a function of $\hat{u}_{\text{rms}}$. As expected, we observe two regimes: (1) for weak and moderate fluctuations, $\hat{u}_{\text{rms}} \lesssim 1$, the gain in height follows a rapidly increasing trend ; (2) for strong fluctuations, $\hat{u}_{\text{rms}} \gtrsim 1$, gains still increase but more slowly. Since the ascended height depends on the flow speed, Fig. 1.3B also shows the soaring efficiency χ, defined as the difference between $\Delta h(\hat{u}_{\text{rms}})$ and $\Delta h(0)$ divided by $w_{\text{rms}}\Delta T$, where w_{rms} is the rms vertical speed of the flow and $\Delta T = 150s$ is the duration of a trial (see Appendix A). If the glider did not attempt to selectively find upward currents, χ would vanish, while $\chi = 1$ corresponds to a glider perfectly capturing vertical currents. As the flow speed increases, the efficiency shows a downward trend that reflects the increasing difficulty in control due to higher levels of fluctuations.

The performance of different gliders soaring simultaneously within the same flow does not vary significantly, indicating that an ensemble of gliders learn a uniquely optimal policy. The performance over different realizations for a single glider varies wildly, with a standard deviation of the final height of the same magnitude as the final height itself when $\hat{u}_{\text{rms}} \approx 1$. Despite this

wide variation, the number of failures, i.e. the glider touches the ground, always decreases rapidly to almost zero with the number of training trials.

1.2.4 Role of wind acceleration and torques

Our learning procedure allows us to test the possible local sensorimotor cues that give good soaring performance. For each cue, we define a mean level and upper and lower thresholds symmetrically around the mean value. The performance was found to be largely independent of the chosen thresholds.

In Fig. 1.3C, we show a comparison between the performance of a few different combinations of the cues. We found that the pairing of vertical wind acceleration and torques, gauged in terms of the average height ascended per trial, works best (results in Fig. 1.3A,B are obtained using this pair). Intuitively, the combination of vertical wind acceleration and torques provides information on the gradient of the vertical wind velocity in two complementary directions, thus allowing the glider to decide between turning or continuing along the same path. Conversely, the vertical wind velocity does indicate the strength of a thermal but it does not guide the glider to the core of the thermal. The pair acceleration and torque allows the glider to climb the thermal towards the core and also detect the edge of a thermal so that the glider can stay within the core. The resulting pattern within a thermal is a spiral that occurs solely from actions based on local observables and minimal memory usage. Temperature fails to improve performance, which could be intuited as the temperature field is highly intermittent and is itself a convoluted function of the turbulent velocity [27, 28].

1.2.5 Control over the angle of attack

Fig. 1.3C shows that control over the angle of attack does not influence significantly the performance in climbing an individual thermal. The angle of attack should play an important

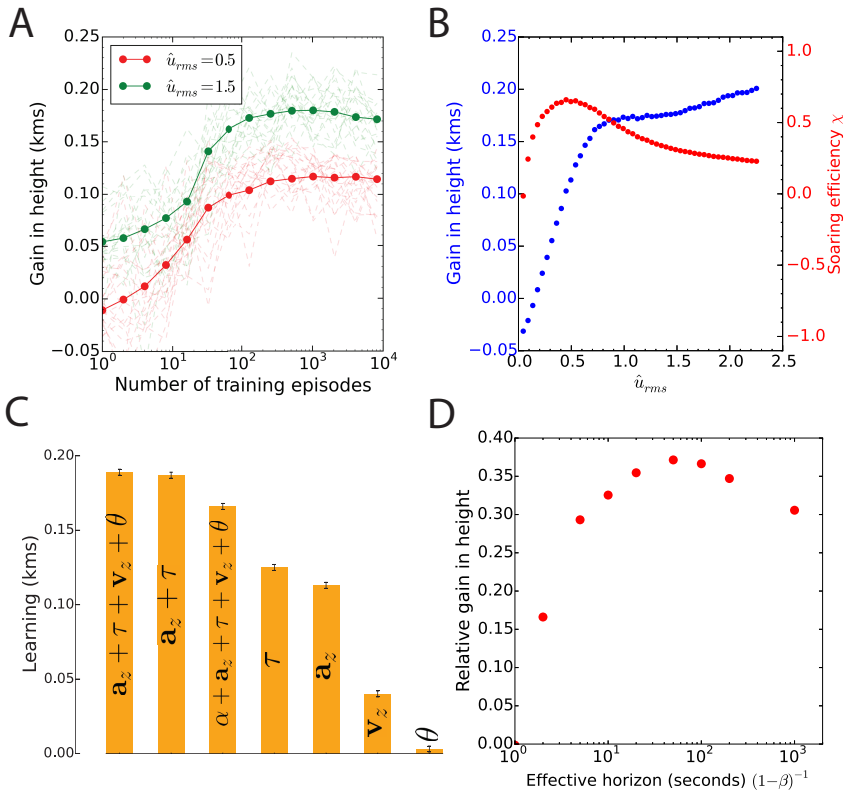

Figure 1.3: (A) The learning curve for two different turbulent fluctuation levels. (B) The average height ascended for different $\hat{u}_{\text{rms}}$ and the soaring efficiency $\chi(\hat{u}_{\text{rms}})$. (C) The average gain in height for different sensorimotor cues. (D) The improvement in height gained w.r.t a greedy strategy with $\beta = 0$.

role though in other situations, namely during cross-country races or bird migration, where gliders need to cover large horizontal distances and control over the horizontal speed and sink rate is needed [11, 29, 30]. To verify this expectation, we considered a simple test case of a glider traversing, without turning, a two-dimensional track consisting of a series of ascending or descending columns of air with turbulence added on top. We found that control over the angle of attack indeed improves the gain in height (see SI) and the glider learns to increase its pace during phases of descent while slowing down during periods of ascending currents. We expect that the differing roles of the angle of attack for soaring between and within thermals holds true for birds as well, a prediction that can be tested in field experiments.

In the sequel we shall analyze the soaring in a single thermal. We fix then for simplicity

the angle of attack at $\sim 9^\circ$ (where the climb rate is the largest, see Fig. 1.1B), and the pair acceleration-torque as sensorimotor cues sensed by the glider, see Fig. 1.3C.

1.2.6 Dependence on the temporal discounting

The performance of the glider as a function of the temporal discount factor β is shown in Fig. 1.3D. The gain in height increases as the effective time horizon $(1-\beta)^{-1}$ grows, reaches a maximum at ≈ 100 seconds and then slowly declines. The best time horizon is comparable with the time-scale of the flow patterns at the height reached by the glider. This demonstrates that long-term planning is crucial for soaring and the importance of a relatively long-term strategy to effectively utilize the ascending thermals.

1.2.7 Optimal flight policy

The Q-function learned by the SARSA algorithm defines the optimal state-action policy via Eq. (1.6). An optimal policy associates the choice of an action to the pair acceleration-torque $(\boldsymbol{a}_z, \tau)$. The optimal action is chosen among the three options : (i) increase the bank angle μ by 5° ; (ii) decrease μ by 5° ; (iii) keep μ unchanged. In Fig. 1.4A, we show a comparison between the policy for the two regimes of weak and strong fluctuations.

The policies in Fig. 1.4 have a few intuitive features that are preserved at different flow speeds. For instance, when the glider experiences a negative wind acceleration, the optimal action is to sharply bank towards the side of the wing which experiences larger lift. When the glider experiences a large positive acceleration and no torque, the glider continues flying along its current path. Despite these similarities, the policies exhibit marked differences, which we proceed to analyze.

For each $\boldsymbol{a}_z, \tau$ pair, it is useful to consider its preferred angles (the green circles in Fig. 1.4), i.e. those angles that the policy leads to if the pair $\boldsymbol{a}_z, \tau$ is maintained fixed. We observe that the

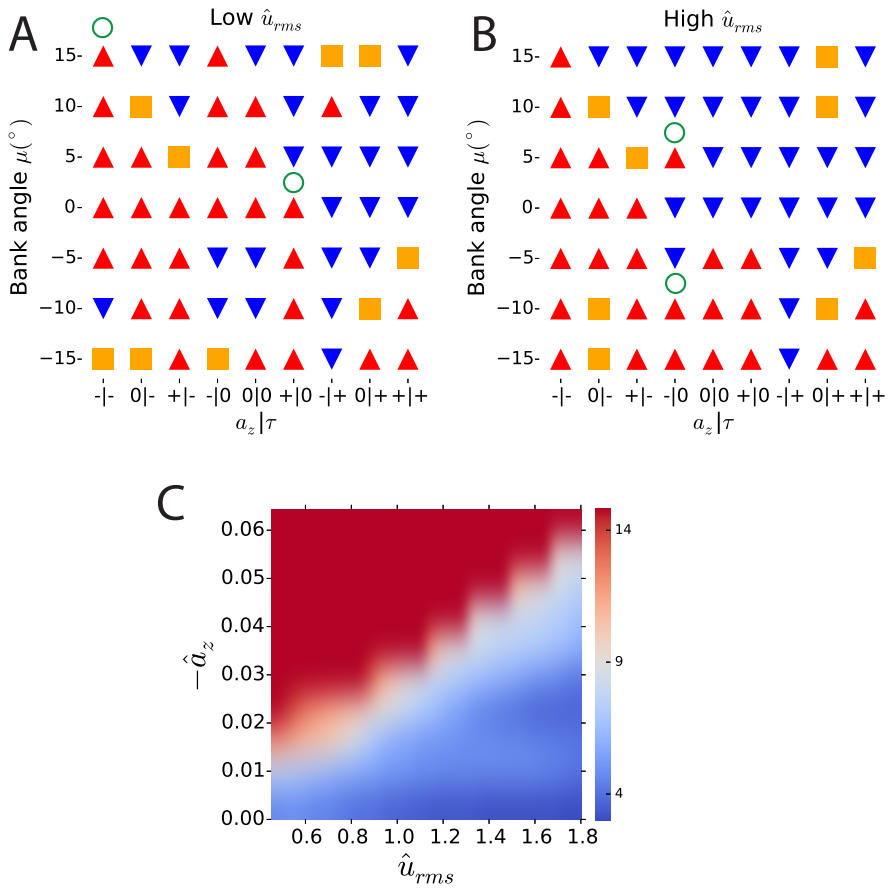

Figure 1.4: Policies of flight for different levels of turbulent fluctuations. (A) $\hat{u}_{\text{rms}} = 0.5$ and (B) $\hat{u}_{\text{rms}} = 1.5$. (C) : a heat map showing the optimal bank angle (see Eq. (1.8)) at a particular $\hat{u}_{\text{rms}}$ and $\hat{a}_z$ with $\tau < 0$.

preferred bank angles of gliders trained in a strong flow are relatively moderate, and the policy in general is more conservative. Consider, for instance, the case of zero torque and zero acceleration (column 5 of the policies in Fig. 1.4). The optimal bank action in the weak flow regime is to turn as much as possible, in contrast to the policy in the strong flow regime, which is to not turn. Another interesting qualitative difference is when the glider experiences negative acceleration and significant torque on the right wing (column 1 of the policies in Fig. 1.4). In the weak flow regime, if the glider is already banked to the left (negative bank angles), the policy is to bank

further left in order to complete a full circle. In the strong flow regime, the policy is once again more conservative, preferring to not risk the full turn.

A policy becoming more conservative and risk-averse as fluctuations increase is consistent with the balance of exploration and exploitation [12]. In a noisy environment, where a wrong decision can lead to highly negative consequences, we expect an active agent to play safe and tend to gather more information before taking action. In a turbulent environment, we expect the glider to exploit (avoid) only significantly large positive (negative) fluctuations along its trajectory while filtering out transient, small-scale fluctuations. In the next subsection, we shall further confirm this expectation by tracking the changes in the optimal policy with the flow speed and extracting a few general principles of the optimal flight policy.

1.2.8 Optimal bank angles

To quantify the description of the optimal policy shown in Fig. 1.4A, we consider the distributions of the bank angle μ given the acceleration $\boldsymbol{a}_z$ and torque τ in the previous time step i.e., $\Pr(\mu^{(t+1)}|\boldsymbol{a}_z^{(t)},\tau^{(t)})$. We define the optimal bank angle as

$$\mu_{\text{opt}}(\boldsymbol{a}_z,\tau) = \arg\max_{\mu^{t+1}} \Pr(\mu^{t+1}|\boldsymbol{a}_z^t,\tau^t), \tag{1.8}$$

and we are interested in the variations of the optimal bank angle with the turbulence level $\hat{u}_{\text{rms}}$. We use a bicubic spline interpolation to smooth the probability distributions and thereby obtain smoothened values for μ_{opt}.

To create a higher resolution in $\boldsymbol{a}_z$, we expand our state space by creating finer divisions in the vertical wind accelerations. Note that the performance with an expanded state space is not significantly better than the one with just three states. Fig. 1.4 shows a heat map of the optimal bank angles at different $\boldsymbol{a}_z < 0$ and $\tau < 0$. For every $\boldsymbol{a}_z$, μ_{opt} drops from the maximum value of 15° to a value closer to zero as $\hat{u}_{\text{rms}}$ increases. Note that since $\tau < 0$, the optimal angles are

biased towards being positive. We define a threshold on the optimal bank angles at 12.5°, which empirically corresponds to the point where the optimal bank angles drop most rapidly as $\hat{u}_{\mathrm{rms}}$ increases. Above (below) the threshold, the angles are considered "high" ("low"). The threshold on the optimal bank angle defined a cutoff on $-\boldsymbol{a}_z$ and thereby an effective "fluctuation filter".

We interpret the fluctuation filter above as follows: At a particular $\hat{u}_{\mathrm{rms}}$, if the glider encounters a fluctuation with $-\boldsymbol{a}_z$ above the cutoff, the glider interprets the fluctuation as significant, i.e. as the large-scale downwards branch of a convective cell, and banks away. Conversely, fluctuations below the cutoff are ignored. In other words, the cutoff defined above gives the level which identifies significantly large fluctuations that require action. Similar behaviors are obtained for $(\boldsymbol{a}_z < 0, \tau = 0)$ and $\tau > 0$ is symmetric with respect to the case $\tau < 0$ just discussed. Conversely, for $\boldsymbol{a}_z > 0$, the glider maintains a bank angle close to zero unless it experiences an exceptionally large torque. These simple principles are the key for effective soaring in fluctuating turbulent environments.

1.3 Discussion

We have shown that reinforcement learning methods cope with strong turbulent fluctuations and identify effective policies of navigation in turbulent flow. Previous works neglected turbulence, which is an essential and unavoidable feature of natural flow. The learned policies dramatically improve the gain of height and the rapidity of climbing within thermals, even when turbulent fluctuations are strong and the glider has reduced control due to its being transported by the flow.

We deliberately kept simple the sensorimotor cues that the glider can sense to guide its flight. In particular, possible cues were local in space and time for two reasons : keep the closest contact with what birds are likely to sense and minimize the mechanical instrumentation needed for the control of autonomously flying vehicles. In the same spirit, we kept simple the

parametrization of the learned policies, by using a relatively coarse discretization of the space of states and actions.

Turbulence has indeed a major impact upon the policy of flight. We explicitly presented how the learned policies of flight modify as the level of turbulence increases. In particular, we quantified the increase of the threshold on the cues needed for the glider to change its parameters of control. We also discussed the simple principles that the policy follows in order to filter out transient, small-scale turbulent fluctuations, and identify the level of the sensorimotor cues which requires actions that modify the parameters of flight of the glider.

We found that the bank angle of the glider is the main control for navigation within a single thermal, which is the main interest of the current work. However, we also considered a very simplified setting mimicking the flight between multiple thermals and there we found that control of the angle of attack is important. Inter-thermals flight is of major interest for birds' migration and glider pilots. MacCready [29] determined the optimal speed to maximize the average cross-country speed as a function of the glider's rate of sink and the velocity of ascent within the thermals. The resulting instrument (the so-called MacCready speed ring) is commonly used by glider pilots with various supplementary empirical prescriptions, which typically tend to be risk-averse. MacCready's prediction was also recently compared to the behavior of various birds [30] along their thermal-dense migratory routes. Their behavior was found to differ from the prediction, viz. a more conservative policy was observed, with slower but less sinking paths that reduce the probability of dramatic losses of height. One possible cause for more conservative policies relates to the uncertainties on the location and the velocity of ascent within the thermals, which was previously considered in the literature [31]. Another possible reason suggested by our results is turbulence along the inter-thermal paths, which is neglected in MacCready's and subsequent arguments. Our methodology can be adapted to realistically model inter-thermal conditions and future work will assess the role of turbulence in the policy of inter-thermal flight.

We identified torque and vertical accelerations as the local sensorimotor cues that most

effectively guide turbulent navigation. Temperature was specifically shown to yield minor gains. The robustness of our results with respect to the modeling of turbulence strongly suggests that the conclusion apply to natural conditions; a sensor of temperature could then be safely spared in the instrumentation for autonomous flying vehicles. More generally, it will be of major interest to implement our predicted policy on remotely controlled gliders and test their flight performance in field experiments. Thanks to our choices discussed above, the mechanical instrumentation needed for control is minimal and can be hosted on commercial gliders without perturbing their aerodynamics. Finally, our flight policy and the nature of the sensorimotor cues that we identified, provide predictions that can be compared with the behavior of soaring birds and could shed light on the decision processes that enable them to perform their soaring feats.

Chapter 1, in full, is a reprint of the material as it appears in Reddy G., Celani A., Sejnowski T. J. & Vergassola M., Learning to soar in turbulent environments, *Proc. Natl. Acad. Sci.*, 113-33:4877-4884, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Glider soaring through reinforcement learning in the field

Soaring birds often rely on ascending thermal plumes in the atmosphere as they search for prey or migrate across large distances [4, 33, 34, 2]. The landscape of convective currents is rugged and rapidly shifts on timescales of a few minutes as thermals constantly form, disintegrate, or are transported away by the wind [6, 7]. How soaring birds find and navigate thermals within this complex landscape is unknown. Reinforcement learning [12] provides an appropriate framework to identify an effective navigational strategy as a sequence of decisions taken in response to environmental cues. Here, we use reinforcement learning to train gliders in the field to autonomously navigate atmospheric thermals. Gliders of two-meter wingspan were equipped with a flight controller that enables an on-board implementation of autonomous flight policies via precise control over their bank angle and pitch. A navigational strategy was determined solely from the gliders' pooled experiences collected over several days in the field using exploratory behavioral policies. The strategy relies on novel on-board methods to accurately estimate the local vertical wind accelerations and the roll-wise torques on the glider, which serve as navigational cues. We establish the validity of our learned flight policy through field experiments, numerical

simulations, and estimates of the noise in measurements that is unavoidably present due to atmospheric turbulence. This is a novel instance of learning a navigational task in the field, where learning is severely challenged by a multitude of physical effects and the unpredictability of the natural environment. Our results highlight the role of vertical wind accelerations and roll-wise torques as viable biological mechanosensory cues for soaring birds, and provide a navigational strategy that is directly applicable to the development of autonomous soaring vehicles.

2.1 Introduction

In reinforcement learning, an animal maximizes its long-term reward by taking actions in response to its external environment and internal state. Learning occurs by reinforcing behavior based on feedback from past experiences. Similar ideas have been used to develop intelligent agents, reaching spectacular performance in strategic games like backgammon [25] and Go [35], visual-based video game play [36] and robotics [37, 38]. In the field, physical constraints fundamentally prevent learning agents from using data-intensive learning algorithms and the optimization of model design needed for quicker learning, which are the conditions most often faced by living organisms.

A striking example in nature is provided by thermal soaring, where the extent of atmospheric convection is not consistent across days and, even under suitable conditions, the locations, sizes, durations and strengths of nearby thermals are unpredictable. As a result, the statistics of training samples are skewed on any particular day. At smaller spatial and temporal scales, fluctuations in wind velocities are due to turbulent eddies lasting a few seconds that may mask or falsely enhance a glider's estimate of its mean climb rate. Further, the measurement of navigational cues using standard instrumentation may be consistently biased by aerodynamic effects, which requires precise quantification. Here, we demonstrate that reinforcement learning can meet the challenge of learning to effectively soar in atmospheric turbulent environments. To

contrast with past work, the maneuvering of an autonomous helicopter in ref. [37] is a control problem that is decoupled from environmental fluctuations and has little trial-to-trial variability. Past autonomous soaring algorithms have largely relied on locating the centroid of a drifting Gaussian thermal [5, 39, 40, 11], which is unrealistic, or have applied learning methods in highly simplified simulated settings [15, 13, 41].

Using the reinforcement learning framework [12], we may describe the behavior of the glider as an agent traversing different states (s) by taking actions (a) while receiving a local reward (r). The goal is to find a behavioral policy that maximizes the value, i.e., the mean sum of future rewards up to a specified horizon. We seek a model-free approach, which estimates the value of different actions at a particular state (called the Q function) solely through the agent's experiences during repeated instances of the task, thereby bypassing the modeling of complex atmospheric physics and aerodynamics (see Methods). The optimal policy is subsequently derived by taking actions with the highest Q value at each state, where the state includes sensorimotor cues and the glider's aerodynamic state.

To identify mechanosensory cues that could guide soaring, we recently combined above ideas with simulations of virtual gliders in numerically generated turbulent flow [42] (Chapter 1). Two cues emerged from our screening: (1) the vertical wind acceleration ($\boldsymbol{a}_z$) along the gliders path; (2) the spatial gradients in the vertical wind velocity across the wings of the glider (ω). Intuitively, the two cues correspond to the gradient of the vertical wind velocity in the longitudinal and lateral directions of the glider, which locally orient it towards regions of higher lift. Simulations described in Chapter 1 further showed that the glider's bank angle is the crucial aerodynamic control variable; additional variables, such as the angle of attack, or other mechanosensory cues, such as temperature or vertical velocity, offer minor improvements when navigating within a thermal.

Below, I describe the methodology to train a glider to autonomously soar in the field. The results from the experiments are described thereafter.

2.2 Methods

2.2.1 Experimental setup

A Parkzone Radian Pro fixed-wing plane of 2-meter wingspan was equipped with an on-board Pixfalcon autonomous flight controller operating on custom-modified Arduplane firmware [47]. The instrumentation available to the flight controller includes a GPS, compass, barometer, airspeed sensor and an inertial measurement unit (IMU). Measurements from multiple instruments are combined by an Extended Kalman Filter (EKF) to give an estimate of relevant quantities such as the altitude z, the sink rate w.r.t ground $\boldsymbol{u}_z$, pitch ϕ, bank angle μ and the airspeed V, at a rate of 10 Hz (see Figure B.2 for the definitions of the angles). Throughout the paper, we use $\mu > 0$ when the plane is banked to the right and $\phi > 0$ for the airplane pitched nose above the horizontal plane. For a given desired pitch ϕ_d and desired bank angle μ_d, the controller modulates the aileron and elevator control surfaces at 400 Hz using a proportional-integral-derivative (PID) feedback control mechanism at a user-set time scale τ (see Table B.1 for parameter values) such that:

$$\tau \frac{d\phi}{dt} = \phi_d - \phi, \tag{2.1}$$

$$\tau \frac{d\mu}{dt} = \mu_d - \mu. \tag{2.2}$$

ϕ_d is fixed during flight and can be used to indirectly modulate the angle of attack, α, which determines the airspeed and sink rate w.r.t air of the glider ($\boldsymbol{v}_z$). Actions of increasing, decreasing or keeping the same bank angle are taken in time steps of t_a by changing the desired bank angle, μ_d, such that μ increases linearly from μ_i to μ_f in time interval t_a:

$$\mu_d(t) = \mu_i + (\mu_f - \mu_i)(t + \tau)/t_a \tag{2.3}$$

2.2.2 Estimation of the vertical wind acceleration

The vertical wind acceleration az is defined as:

$$\boldsymbol{a}_z \equiv \frac{d\boldsymbol{w}_z}{dt} = \frac{d}{dt}(\boldsymbol{u}_z - \boldsymbol{v}_z) \tag{2.4}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the velocities of the glider w.r.t the ground and air respectively, and $\boldsymbol{w}$ is the wind velocity. Here, we have used the relation $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}$. An estimate of $\boldsymbol{u}$ is obtained in a straightforward manner from the EKF, which combines the GPS and barometer readings to form the estimate. However, $\boldsymbol{v}_z$ is confounded by various aerodynamic effects that significantly affect it on time scales of a few seconds (Figure B.3). Artificial accelerations introduced due to these effects impair accurate estimation of the wind acceleration and thus alter the perceived state during decision-making and learning. Two effects significantly influence variations in $\boldsymbol{v}_z$: (1) Sustained pitch oscillations with a period of a few seconds and varying amplitude, and (2) Angle of attack variations, which occur in order to compensate for the imbalance of lift and weight while rolling. In Appendix B, we present a detailed analysis of the longitudinal motions that affect the glider, which is summarized here for conciseness. Changes in $\boldsymbol{v}_z$ can be approximated as:

$$\Delta\boldsymbol{v}_z = -V(\Delta\alpha - \Delta\phi) \tag{2.5}$$

where the Δ denotes the deviation from their value during steady, level flight. We obtain $\Delta\phi$ directly from on-board measurements whereas $\Delta\alpha$ can be approximated for bank angle μ as:

$$\Delta\alpha = \frac{\alpha_0 - \alpha_i}{1/\cos\mu - 1} \tag{2.6}$$

where α_0 is the angle of attack at steady, level flight and α_i is a parameter which depends on the geometry and the angle of incidence of the wing. The constant pre-factor $(\alpha_0 - \alpha_i)$ is inferred from experiments. Measurements of $\boldsymbol{u}_z$ together with the estimate of $\Delta\boldsymbol{v}_z$ are now used to estimate

the vertical wind velocity $\boldsymbol{w}_z$ up to a constant term, which can be ignored as it does not affect $\boldsymbol{a}_z$. The vertical wind acceleration $\boldsymbol{a}_z$ is then obtained by taking the derivative of $\boldsymbol{w}_z$ and is further smoothed using an exponential smoothing kernel of time scale σ_a (Figure B.4).

2.2.3 Estimation of the vertical wind velocity gradients across the wings

Spatial gradients in the vertical wind velocity induce a roll-wise torque on the plane, which we estimate using the deviation of the measured bank angle from the expected bank angle. The total roll-wise torque on the plane has contributions from three sources (1) the feedback control of the plane, (2) spatial gradients in the wind including turbulent fluctuations, and (3) roll-wise moments created due to various aerodynamic effects. Here, we follow an empirical approach: we note that the latter two contributions perturb the evolution of the bank angle from equation (2.2). We can then write an effective equation,

$$\frac{d\mu}{dt} = (\mu_d - \mu)/\tau + \omega(t) + \omega_{\text{aero}}(t), \tag{2.7}$$

where $\omega(t)$ and $\omega_{\text{aero}}(t)$ are contributions to the roll-wise angular velocity due to the wind and aerodynamic effects respectively. We empirically find four major contributions to $\omega_{\text{aero}}(t)$: (1) the dihedral effect, which is a stabilizing moment due to the effects of sideslip on a dihedral wing geometry, (2) the over-banking effect, which is a destabilizing moment that occurs during turns with small radii, (3) trim effects, which create a constant moment due to asymmetric lift on the two wings, and (4) a loss of rolling moment generated by the ailerons when rolling at low airspeeds. We quantify the contributions from the four effects and model their dependence on the bank angle (see Appendix B for more details on modeling and calibration). A estimate of ω is then obtained as:

$$\omega = \frac{d\mu}{dt} - (\mu_d - \mu)/\tau - \omega_{\text{aero}}(t). \tag{2.8}$$

Finally, an exponential smoothing kernel is applied to obtain a smoothed ω (Figure B.5).

2.2.4 Design of the learning module

The navigational component of the glider is modeled as a Markov Decision Process (MDP), closely following the implementation used in Chapter 1. The Markovian transitions are discretized in time into intervals of size t_a. The state space consists of the possible values taken by $\boldsymbol{a}_z$, ω and μ. To make the learning feasible within experimental constraints and to maintain interpretability, we use a simple tile coding scheme to discretize our state space: continuous values of $\boldsymbol{a}_z$ and ω are each discretized into three states $(+,0,-)$, partitioned by thresholds $\pm K_a$, $\pm K_\omega$ respectively. The thresholds are set at ± 0.8 times the standard deviation of $\boldsymbol{a}_z$ and ω. Since the width of the distributions of $\boldsymbol{a}_z$ and ω can vary across days, the data obtained on a particular day is normalized by the standard deviation calculated for that day. In effect, the filtration threshold to detect a signal against turbulent noise is higher on days with more turbulence. The consequence is that the behavior of the learned strategy could change across days, adapting to the recent statistics of the environment. The bank angle takes five possible values $0^\circ, \pm 15^\circ, \pm 30^\circ$, while the three possible actions allow for increasing, decreasing by 15° or keeping the same bank angle. In summary, we have a total of $3 \times 3 \times 5 = 45$ states in the state space and 3 actions in the action space.

We choose the local vertical wind acceleration $\boldsymbol{a}_z$ obtained in the next time step as the reward function. The choice of $\boldsymbol{a}_z$ as an appropriate reward signal is motivated by observations made in simulations from Chapter 1. In Appendix B, we show that the obtained policy using $\boldsymbol{a}_z$ as the reward function is equivalent to a policy that also maximizes the expected gain in height.

2.2.5 Learning the thermalling strategy in the field

Data collected in the field is split into (s, a, s', r) quadruplets containing the current state s, the current action a, the next state s' and the obtained reward r, which are pooled together to obtain the transition matrix $T(s'|s,a)$ and reward function $R(s,a)$. Value iteration methods are used to estimate the Q values from T and R. The learning process is offline and off-policy; specifically, we begin training with a random policy that takes the three possible actions with equal probability irrespective of the current state as our behavioral policy, which was used for 12 out of the 15 days of training. For the other days, a softmax policy [12] with temperature set to 0.3 was used. For softmax training, the Q values were first estimated from the data obtained in the previous days and then normalized by the difference between the maximum and minimum Q values over the three possible actions at a particular state, as described in Chapter 1.

Using a fixed, random policy as our behavioral policy slows learning as state-action pairs that rarely appear in the final policy are still sampled. On the other hand, calibrating the parameters necessary for the unbiased measurement of $\boldsymbol{a}_z$ and ω (see Appendix B) is performed simultaneously with learning, which considerably reduces the number of days required in the field. Importantly, offline learning permits us to continuously monitor the variance of the estimated Q values by bootstrapping from the set E of accumulated (s, a, s, r) quadruplets up to a particular point. Specifically, $|E|$ samples are drawn with replacement from E and Q values are obtained for each state-action pair via value iteration. The steps are repeated and the average of the bootstrapped standard deviations in Q over all the state-action pairs is used as a measure of learning progress, as shown in Figure 2.2A.

We expect certain symmetries in the transition matrix and the reward function, which we exploit in order to expedite our learning process. Particularly, we note that the MDP is invariant to an inversion of sign in the bank angle $\mu \to -\mu$. This transforms a state as $(\boldsymbol{a}_z, \omega, \mu) \to (\boldsymbol{a}_z, -\omega, -\mu)$ and inverts the action from that of increasing the bank angle to decreasing the bank

angle and vice-versa. We symmetrize T and R as

$$T_{\text{sym}} = (T^{+} + T^{-})/2, \tag{2.9}$$

$$R_{\text{sym}} = (R^{+} + R^{-})/2, \tag{2.10}$$

where $+$ and $-$ denote the obtained values and those computed by applying the inverting transformation respectively. Finally, T_{sym} and R_{sym} are used to obtain a symmetrized Q function, which results in a symmetric policy as shown in Figure 2.2B. To conveniently obtain the policy that uses only $\boldsymbol{a}_z$ (Figure 2.3D), the above procedure is repeated with the threshold for ω (K_ω) set to infinity.

2.2.6 Testing the performance of the learned policy in the field

To obtain the data shown in Figure 2.3B, the glider is first sent autonomously to an arbitrary but fixed location 250 m above ground level. The learned thermalling policy is then turned on and the mean climb rate i.e., the total height gained divided by the total time, is measured over a 3-minute interval. To obtain the control data, the glider instead follows a random policy, which takes the three possible actions with equal probability. The trials where we observe little to no atmospheric convection were filtered out by imposing a threshold on the standard deviation of the vertical wind velocity over the 3-minute trial. In Figure B.6, we show the distribution of the standard deviation in $\boldsymbol{w}_z$ collected from 240 3-minute trials over 9 days. Trials below the threshold chosen as the 25th percentile mark (red, dashed line) are not used for our analysis.

2.2.7 Testing the performance for different wingspans in simulations

Soaring performance is analyzed in simulations similar to those described in Chapter 1 and adapted to reflect the constraints faced by our glider and the environments typically observed in the field.

The atmospheric model consists of two components: (1) a kinematic model of turbulence that reproduces the statistics of wind velocity fluctuations in the convective atmospheric boundary layer, and (2) the positions, sizes and strengths of updrafts and downdrafts. The temporal and spatial statistics of the generated velocity field satisfy the Kolmogorov and Richardson laws [10] and the mean velocity profile in the convective boundary layer [6], as described in Appendix B. Stationary updrafts and downdrafts of Gaussian shape are placed on a staggered lattice of spacing $\sim$125m on top of the fluctuating velocity field. Specifically, their contribution to the vertical wind velocity at position r is given by

$$w_z = \pm W e^{-(r_\perp - r_\perp^0)^2/2R^2}, \tag{2.11}$$

where $r_\perp^0$ is the location of the center of the up(down)draft in the horizontal plane, W is its strength and R is its radius. W is drawn from a half-normal distribution of scale 1.5m/s whereas the radius is drawn from a (positive) normal distribution of mean 40m and deviation 10m. Gaussian white noise of magnitude 0.2m/s is added as additional measurement noise.

We assume the glider is in mechanical equilibrium; the lift, drag and weight forces on the glider are balanced, except for centripetal forces while turning. The parameters corresponding to the lift and drag curves and the (fixed) angle of attack are set such that the airspeed is V = 8m/s and the sink rate is 0.9m/s at zero bank angle, which match those measured for our glider in the field. Control over bank angle is similar to those imposed in the experiments i.e., the bank angle switches linearly between the angles $0^\circ, \pm 15^\circ, \pm 30^\circ$ in a time interval t_a, corresponding to the time step between actions. The gliders trajectory and wind velocity readings are updated every 0.1s. The vertical wind acceleration is derived assuming that the glider directly reads the local vertical wind velocity. The vertical wind velocity gradients across the wings are estimated as the difference between the vertical wind velocities at the two ends of the wings. The readings are smoothed using exponential smoothing kernels; the smoothing parameters in experiments are

chosen to coincide with those that yield the most gain in height in simulations.

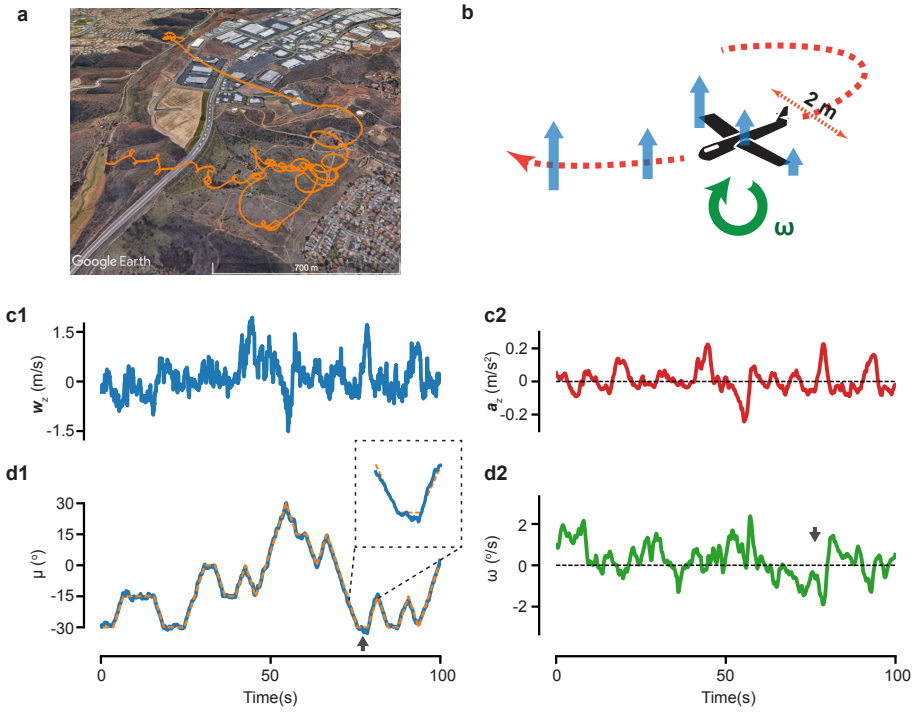

Figure 2.1: (a) A trajectory of our glider soaring in Poway, California. (b) A cartoon of the glider showing the available navigational cues. (c) A sample trace of the estimated vertical wind velocity $\boldsymbol{w}_z$ and $\boldsymbol{a}_z$ obtained in the field. (d) The measured bank angle μ and the estimated ω during the same trial as in (c).

2.3 Results

To learn to soar in the field, we used a glider (of two-meter wingspan) with autonomous soaring capabilities (Figure 2.1A-B). The glider is equipped with a flight controller, which implements a feedback control system used to modulate the glider's ailerons and elevator such that a desired bank angle and pitch are maintained. Relevant measurements, such as the altitude, ground velocity ($\boldsymbol{u}$), airspeed, bank angle (μ) and pitch, are made continuously at 10 Hz using standard instrumentation (see Methods). At fixed time intervals, the glider changes its heading by modulating its bank angle in accordance with the implemented behavioral policy.

Noise and biases that affect learning in the field require the development of appropriate methods to extract environmental cues from sensory devices measurements. We found that estimating a_z by the derivative of the vertical ground velocity (u_z), is significantly biased by longitudinal motions of the glider about the pitch axis as the glider responds to an imbalance of forces and moments while turning. By modeling the glider's longitudinal dynamics, we obtain an unbiased estimate of the local vertical wind velocity (w_z), and a_z as its derivative (Methods). The estimation of the spatial gradients across the wings, ω, poses a greater challenge as it involves the difference between two noisy measurements at relatively close positions. The key observation we used here is that the glider rolls due to contributions from vertical wind velocity gradients, the feedback control mechanism and various aerodynamic effects. The resulting roll-wise torque can be estimated from the small deviations of the true bank angle from the desired one, and a novel dynamical model allows us to separate the ω contribution due to velocity gradients from the other effects (Methods). A sample trace of the resulting unbiased estimate of ω is shown in Figure 2.1C-D, together with traces of the vertical wind velocity, w_z, μ and unbiased estimates of a_z.

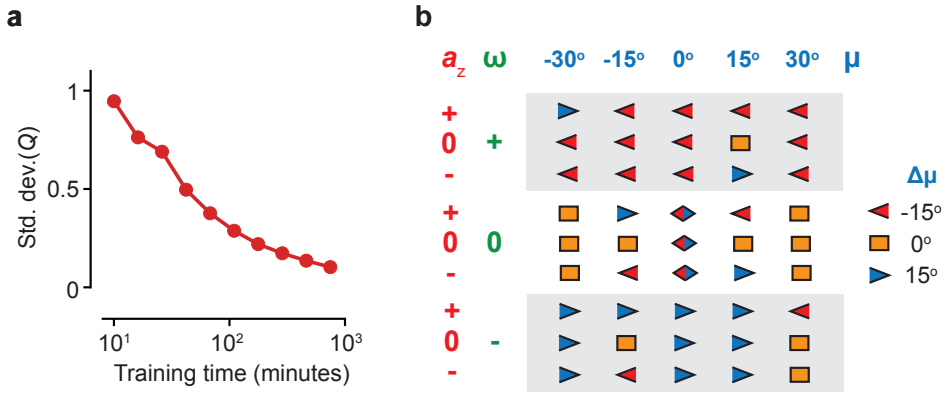

Figure 2.2: (a) The convergence of Q values during learning as measured by the standard deviation of the mean Q value vs training time in the field. (b) The final learned policy. Each symbol corresponds to the best action (increasing/decreasing the bank angle μ by 15° or maintain the same μ)

Equipped with a proper procedure for estimating environmental cues, we next addressed the specifics of learning in the field. First, to constrain our state space, we discretized the range

of values of $\boldsymbol{a}_z$ and ω into three states each, positive high (+), neutral (0) and negative high (-). Second, we found that learning is accelerated by choosing $\boldsymbol{a}_z$ attained at the subsequent time step as the reward signal. The choice of $\boldsymbol{a}_z$ (rather than $\boldsymbol{w}_z$) is an instance of reward shaping that is justified in Appendix B, where we show that using $\boldsymbol{a}_z$ as a reward still leads to a policy that optimizes the long-term gain in height. This property is a special case of our general result that a particular reward function or its time derivatives (of any order) yield the same optimal policy (Appendix B). Choosing $\boldsymbol{w}_z$ as the reward fails to drive learning in the soaring problem, possibly because the velocities (and thus the rewards) are correlated across states and their temporal statistics strongly deviates from the Markovianity assumption in reinforcement learning methods [12]. Indeed, velocity fluctuations in turbulent flow are long-correlated, i.e. their correlation timescale is determined by the largest timescale of the flow (see for instance Fig. 9 of ref. [43]), which is of the order of minutes in the atmosphere. Conversely, the correlation timescale of accelerations is controlled by the smallest timescale [43, 44, 45] (the dissipation timescale in Fig. 7 of ref. [43]). This is estimated to be only a fraction of a second, which is much smaller than the time interval between successive actions. The previous experimental observations can be rationalized by the combination of the power-law spectrum of turbulent velocity fluctuations in the atmosphere and the extra factor of frequency squared in the spectrum of acceleration vs velocity fluctuations [45]. Finally, the glider's experiences, represented as state-action-state-reward quadruplets, (s_t, a_t, s_{t+1}, r_t), were cumulatively collected (over 15 days) into a set E using explorative behavioral policies. Learning is monitored by bootstrapping the standard deviation of the Q values from E (Figure 2.2A), calculated using value iteration methods (Methods).

The navigational strategy derived at the end of the training period is presented in Figure 2.2B, which shows the actions deemed optimal for the 45 possible states. Remarkably, the rows corresponding to $\omega = 0$ resemble the so-called Reichmann rules [14] – a set of simple heuristics for soaring, which suggest a decrease/increase in bank angle when the climb rate increases/decreases. Our strategy also gives a prescription for bank: for instance, when $\boldsymbol{a}_z$ and ω

are both positive (top row in Figure 2.2B) i.e., in a situation when better lift is available diagonal to the glider's heading, it is advantageous to bank not to the extreme but rather maintain an intermediate value between -30° and -15°. Importantly, the learned leftward/rightward bias in bank angle on encountering a positive/negative torque validates our estimation procedure for ω.

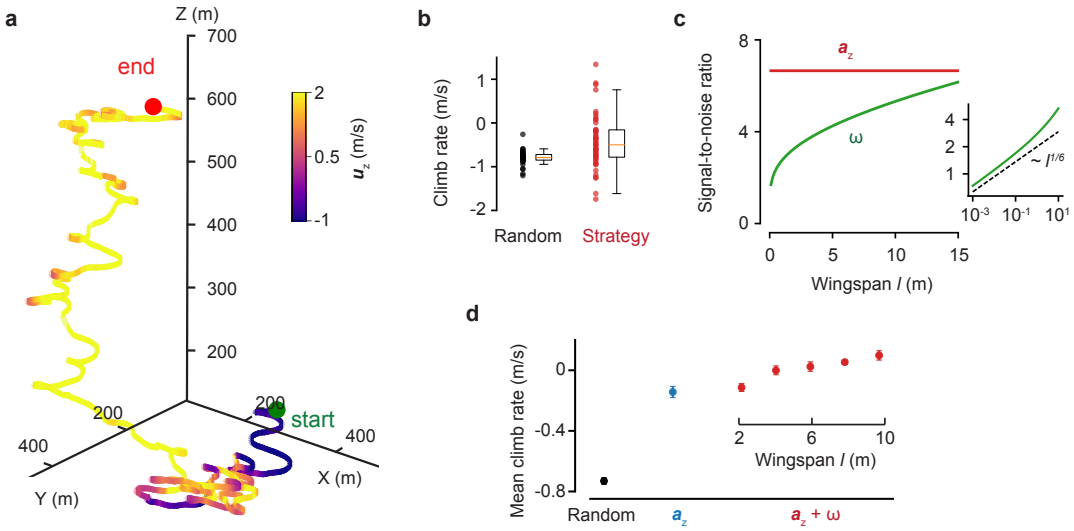

Figure 2.3: (a) A 12-minute trajectory of the glider executing the learned strategy. (b) Measured climb rate of a random policy is compared against the learned strategy over 3-minute trials. (c) SNR in ω and a_z estimation vs wingspan (l). (d) Mean climb rate for different wingspans in simulations.

In Figure 2.3A, we show a sample trajectory of the glider implementing the navigational strategy in the field to remain aloft for 12 minutes while spiraling to the height of low-lying clouds (see also Figure B.1). On a day with strong atmospheric convection, the time spent aloft is limited only by visibility and the receivers range as the glider soars higher or is constantly pushed away by the wind. A significant improvement in median climb rate of 0.35 m/s was measured in the field by performing repeated 3-minute trials over five days (Figure 2.3B, Mann-Whitney $U = 429$, $n_{\text{control}} = 37$, $n_{\text{strategy}} = 49$, $p < 10^{-4}$ two-sided). Notably, this value reflects a general improvement in performance averaged across widely variable conditions without controlling for the availability of nearby thermals.

To examine possible advantages of larger gliders due to improved torque estimation, we

further analyzed soaring performance for different wingspans (l). While the naive expectation is that the signal-to-noise ratio (SNR) in the estimation of ω scales linearly with l, we show that the effects of atmospheric turbulence lead to a much weaker $l^{1/6}$ scaling (Appendix B). Since testing our prediction would require a series of gliders with different wingspans, we turned to numerical simulations of the convective boundary layer, adapted to reflect our experimental setup (Appendix B). Results shown in Figure 2.3C-D are consistent with the predicted scaling. Intuitively, the weak 1/6th exponent arises because the improvement in gradient estimation is offset by the larger turbulent eddies, which only have a sweeping effect for smaller wingspans, and contribute to velocity differences across the wings as l increases. Our calculation yields an estimate of the SNR ~ 4 for typical experimental values; similar arguments for $\boldsymbol{a}_z$ yield an SNR ~ 7. Experimental results, together with simulations and SNR estimates, establish $\boldsymbol{a}_z$ and ω as robust navigational cues for thermal soaring.

2.4 Discussion

The real-world intricacies of soaring impose severe constraints on the complexity of the underlying models, reflecting a fundamental trade-off between learning speed and performance. Notably, the choice of a proper reward signal was crucial to make learning feasible with the limited samples available. Though reward shaping has received some attention in the machine learning community [46], its relevance for behaving animals remains poorly understood. We remark that our navigational strategy constitutes a set of general reactive rules with no learning performed during a particular thermal encounter. A soaring bird may use a model-based approach of constantly updating its estimate of nearby thermals location based on recent experience and visual cues. Still, the importance of vertical wind accelerations and torques for our policy suggests that they are likely useful for any other strategy; our methods to estimate them in a glider suggest that they should be accessible to birds as well. The hypothesis that birds utilize those mechanical

cues while soaring can be tested in experiments.

Finally, we note that single-thermal soaring is just one face of a multifaceted question: how should a migrating bird or a cross-country glider fly among thermals over hundreds of kilometers for a quick, yet risk-averse, journey [29, 30, 31]? Answers to this question, coupled with our current work, pave the way towards a better understanding of how birds migrate and the development of autonomous vehicles that can endlessly fly with minimal energy cost.

Chapter 2, in full, is a reprint of the material as it appears in Reddy G., Wong Ng J., Celani A., Sejnowski T. J. & Vergassola M., Glider soaring via reinforcement learning in the field, *Nature*, Vol. 562, pp. 236-239, 2018. The dissertation author was the primary investigator and author of this paper.

Appendix A

Supplemental Information for Chapter 1

A.1 Modeling the atmospheric boundary layer

Conditions ideal for thermal soaring typically occur during a sunny day, when a strong temperature gradient between the surface of the Earth and the top of the atmospheric boundary layer drive strong convective flow. The soaring flight of birds and gliders primarily occurs within this convective boundary layer. The mechanical and thermal forces within the boundary layer generate turbulence characterized by strongly fluctuating wind velocities. We simulated those turbulent conditions in two different ways: (1) a direct numerical simulation of Rayleigh-Bénard (RB) convection, which captures the basic physical mechanisms of thermal formation, (2) a kinematic model of turbulent fluctuations that reproduces the statistical features of turbulence in the convective boundary layer. The second model accurately captures the Kolmogorov and Richardson laws, and the mean velocity profile of the atmospheric boundary layer. The RB flow allows us to explore the role of temperature as a cue for orientation in turbulent environments.

A.1.1 Rayleigh-Bénard convective flow

Our simulations involve the numerical integration of Navier-Stokes equations with coupled velocity and temperature fields simplified by the Boussinesq approximation. When the Rayleigh number Ra is beyond a critical value $\sim 10^3$, the thermally-generated buoyancy drives instabilities in the flow. In this regime, the flow is characterized by large-scale convective cells and turbulent eddies at every length scale. In the atmosphere, the Rayleigh number can reach up to Ra = $10^{15} - 10^{20}$. In such high Rayleigh number regimes, the flow is strongly turbulent - numerical simulations of convection in the atmosphere are thus plagued by the same limitations of simulating fully developed turbulent flows. We simulated 3D Rayleigh-Bénard convection with a Rayleigh number Ra = 10^8 and a Prandtl number Pr = 0.7 using the Gerris Flow Solver [18]. The floor and the ceiling of the cubical simulation box are no-slip and are fixed at temperatures of unity and zero, respectively. We impose periodic boundary conditions on the side walls. The equations involved are the perturbed velocity ($\boldsymbol{u}$), temperature (θ) field equations about the mean field [19] :

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} = -\nabla P + \left(\frac{\mathrm{Pr}}{\mathrm{Ra}}\right)^{1/2}\nabla^2\boldsymbol{u} + \theta\hat{\boldsymbol{z}}\,, \tag{A.1}$$

$$\frac{\partial \theta}{\partial t} + \boldsymbol{u}\cdot\nabla\theta = \frac{1}{(\mathrm{PrRa})^{1/2}}\nabla^2\theta\,, \tag{A.2}$$

along with the incompressibility condition ($\nabla\cdot\boldsymbol{u} = 0$). For accurate simulations, the grid spacing in the bulk δ_b should be chosen smaller than the Kolmogorov viscous length scale η of the flow. If the side of the cubical box is h, the Kolmogorov scale can be approximated by $\eta/h \approx \pi(\mathrm{Pr}/\mathrm{RaNu})^{1/4}$ [19]. The Nusselt number Nu is defined as the ratio of convective to conductive heat transfer in the flow. At our parameter values, the Nusselt number can be approximated by $\mathrm{Nu} \approx 0.124\mathrm{Ra}^{0.309}$ [19]. We thereby obtain $64\eta/h \approx 0.8$, and thus we use a spacing of $\delta_b/h = 1/64$ within the bulk. The grid spacing is required to be smaller at the no-slip boundaries due to the formation of the thermal and viscous boundary layers; the Grotzbach criterion [32] suggests 3-5 grid points within the boundary layers for accurate numerical simulations. The

thickness of the thermal boundary layer can be approximated by $\delta_T/h \approx 1/2\mathrm{Nu}$ [20] and the viscous boundary layer thickness is $\delta_v = \mathrm{Pr}\delta_T$. This gives $\delta_T/h = 0.016$. We found that using a grid spacing of $h/256$ within the boundary layers is sufficient to ensure stability of the numerical integration scheme and proper resolution of the fields.

In summary, our setup consists of a cubical grid symmetric about the center and the mesh size is : $h/256$ up to a height of $0.025h$, $h/128$ from $0.025h$ to $0.05h$ and finally $h/64$ in the bulk. In Fig. A.1 we show the velocity and temperature profiles of the flow. Also shown is the Nusselt number defined as (see [9]) :

$$\mathrm{Nu} = (\langle \boldsymbol{u}_z\theta\rangle - \kappa\langle\partial_z\theta\rangle)/\kappa\Delta\theta/h = (\mathrm{PrRa})^{1/2}\langle \boldsymbol{u}_z\theta\rangle - \langle\partial_z\theta\rangle\,, \tag{A.3}$$

where $\kappa = (\mathrm{PrRa})^{-1/2}$ is the effective thermal conductivity after rescaling, $\Delta\theta = 1$ is the (rescaled to unity) temperature difference between the hot and cold plates and $h = 1$ is the (rescaled to unity) distance between the plates. The numerically obtained value $\mathrm{Nu} \approx 32$ matches well with previous values in the literature (see Figure 1(a) in Ref. [20]).

A.1.2 A kinematic model of the convective boundary layer

The Atmospheric Boundary Layer (ABL) is the lowest region of the atmosphere and extends up to a height of about 1-2 Km. Above the ABL, the flow is nearly geostrophic i.e., winds flow along isobars due to the balance of pressure gradient forces and the Coriolis force. On a sunny day, the boundary layer is characterized by convective flows and is roughly structured in four layers:

- Surface layer : This is typically a thin layer of a few meters dominated by shear forces. The wind velocity profile has a logarithmic dependence on the height z, i.e. $u(z) \sim \log(z/z_0)$, where z_0 depends on the surface roughness.

- Free convection layer : This is a matching layer between the surface and the mixing layers.

In this layer, the velocity profile features a Kolmogorov scaling $u^2(z) \sim z^{2/3}$. The layer extends from a few meters up to $0.1 z_i$, where z_i is the inversion height, the top of the ABL.

- Mixed layer : In this layer, shear forces are negligible and the surface is irrelevant. Convective mixing forces the velocity profile, temperature and velocity correlation lengths to be uniform with height. The layer extends from $0.1 z_i$ to z_i.

- Inversion layer : The top of the ABL has a capping inversion layer characterized by cold temperatures, strong winds and clouds.

In our simulations, we resolve the free convection layer and the mixed layer with the inversion height $z_i \sim 1$ Km. We use a kinematic model of turbulence that extends the one in [21] to the inhomogeneous case and statistically reproduces the Kolmogorov and Richardson laws and the velocity profile of the atmospheric boundary layer. A Gaussian ascending core is added on top of the turbulent fluctuations and provides a mean, z-independent, ascending flow :

$$\boldsymbol{u}_z^{\text{thermal}} \propto e^{-\frac{(\boldsymbol{r}_\perp - \boldsymbol{r}_\perp^{\text{center}})^2}{2R^2}}, \tag{A.4}$$

where $\boldsymbol{r}_\perp$ is the two-dimensional position vector in the horizontal plane, $\boldsymbol{r}_\perp^{\text{center}}$ is the location of the center of the thermal and R is its radius.

The fluctuating field is a composition of flows of different integral length scales. The velocity at height z has contributions from flows of length scales l_n greater than z :

$$\boldsymbol{u}(\boldsymbol{r}_\perp, z, t) = \sum_{l_n > z} c_n \boldsymbol{u}(\boldsymbol{r}_\perp, z, t | l_n), \tag{A.5}$$

where $\boldsymbol{u}(\boldsymbol{r}_\perp, z, t | l_n)$ is the velocity contribution due to the flow with length scale l_n. Amplitudes are normalized to have $\langle \boldsymbol{u}(\boldsymbol{r}_\perp, l_n, t | l_n) . \boldsymbol{u}(\boldsymbol{r}_\perp, l_n, t | l_n) \rangle = 1$. Hereafter, $\boldsymbol{u}(\boldsymbol{r}_\perp, z, t | l_n)$ will be denoted

by $\boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t)$. We expand in a spatial discrete Fourier transform with spatial frequencies $\boldsymbol{k}$:

$$\boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) = \sum_{\boldsymbol{k}} \hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, k_z, t) e^{i(\boldsymbol{k}_\perp \cdot \boldsymbol{r}_\perp + k_z z)} \,. \tag{A.6}$$

The energy spectrum of velocity fluctuations at relevant scales follows the Kolmogorov law $E(k) \sim k^{-5/3}$, where $k = |\boldsymbol{k}|$. From the Kolmogorov law, we require the spatial energy spectrum $E(k) = 4\pi k^2 \langle |\hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, k_z, t)|^2 \rangle \sim k^{-5/3}$ and we have $\langle |\hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, k_z, t)|^2 \rangle \sim k^{-11/3}$ for $k > 1/l_n$. The time evolution of the real part (and similarly for the independent imaginary part) of each mode is modeled as an Ornstein-Uhlenbeck process

$$d\hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) = -\hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) \left(\frac{dt}{\tau_{nk}} \right) + \sigma_{nk} dW_t \,, \tag{A.7}$$

where dW_t is a Wiener process, τ_{nk} is a time scale that depends on n and k, and σ_{nk} is the amplitude of fluctuations. The equation above ensures that the temporal correlations of each mode decay with time scale τ_{nk}, which obeys the dimensional Kolmogorov scaling $\tau_{nk} \sim (l_n k)^{-2/3}$. At steady state, the energy of mode $\boldsymbol{k}$ equals $\langle |\hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, k_z, t)|^2 \rangle = 3\sigma_{nk}^2 \tau_{nk}$. Since the contribution due to a flow of length scale l_n is required to cut off at height l_n and vanishes at $z = 0$, we impose the supplementary condition that the Fourier expansion of $\boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t)$ has only sinusoidal contributions in the vertical direction. We thereby have two conditions on the Fourier components :

$$\hat{\boldsymbol{u}}_n(-\boldsymbol{k}_\perp, -k_z, t) = \hat{\boldsymbol{u}}_n^*(\boldsymbol{k}_\perp, k_z, t); \quad \hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, -k_z, t) = -\hat{\boldsymbol{u}}_n(\boldsymbol{k}_\perp, k_z, t) \,. \tag{A.8}$$

The first condition enforces that the flow is real and the second one enforces the vanishing at the ground.

The two conditions (A.8) can be used to reorganize the sum in (A.6) and elementary

calculations lead to the following expression for the two-point velocity correlation function :

$$\langle \boldsymbol{u}_n(\boldsymbol{r}_\perp + \boldsymbol{l}_\perp, z, t) \cdot \boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) \rangle = 16 \tilde{\sum} \sin^2(k_z z) \left[\langle \hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) \cdot \hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) \rangle \cos(\boldsymbol{k}_\perp \cdot \boldsymbol{l}_\perp) \right] . \tag{A.9}$$

The sum $\tilde{\sum}$ runs over the set of independent wave vectors (which is restricted by the two conditions (A.8)). The real and imaginary components of the modes are Gaussian and independent of each other. The Kolmogorov scaling of the amplitudes of the modes mentioned above finally gives $\langle \boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) \cdot \boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) \rangle \sim z^{2/3}$. Due to the imposed sinusoidal constraints, the scaling flattens out around $z = l_n/4$.

The extent of the simulation lattice depends on the integral length scale l_n of the flow. The lattice for the nth flow has dimensions $4z_i \times 4z_i \times l_n$ with each dimension discretized into 64 points. The spatial frequencies are of the form $(p/4z_i, q/4z_i, r/l_n)$ where $p, q, r = 0, 1, 2, \ldots, 63$. To integrate the Fourier components, we use the standard standard stochastic Runge-Kutta update rule :

$$\hat{\boldsymbol{u}}_{nk} \to \hat{\boldsymbol{u}}_{nk}(1 - \delta + \frac{1}{2}\delta^2) + a_{nk}\mathcal{N}(0, \sqrt{\delta}) - a_{nk}[\rho \mathcal{N}(0, \sqrt{\delta}) + \sqrt{1 - \rho^2}\mathcal{N}(0, \sqrt{\delta'})] , \tag{A.10}$$

where $a_{nk} = \sigma_{nk}\sqrt{\tau_{nk}}$, $\mathcal{N}(\mu, \sigma)$ is a normal random variable with mean μ and variance σ^2, and the notation $\hat{\boldsymbol{u}}_{nk}$ indicates that the same update rule holds for the real and imaginary parts of the modes with $|\boldsymbol{k}| = k$. At steady state, $\langle \hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) \cdot \hat{\boldsymbol{u}}_{nR}(\boldsymbol{k}_\perp, k_z, t) \rangle = 3a_{nk}^2/2$. Here, $\delta = dt_k/\tau_{nk}$ and $\delta' = \delta^3/(12 - 3\delta^2)$. It follows from (A.9) that

$$\langle \boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) \cdot \boldsymbol{u}_n(\boldsymbol{r}_\perp, z, t) \rangle = 24 \tilde{\sum} \sin^2(k_z z) a_{nk}^2 , \tag{A.11}$$

with $a_{nk}^2 \propto k^{-11/3}$. The length scales are spaced logarithmically, i.e. $l_n = l_1 \lambda^{n-1}$ and the contribution $n = N$ with the largest length scale has $l_N = 4z_i$. By numerically calculating the velocity

magnitudes from (A.11), we scale the value of the coefficients c_n to obtain the behavior which is appropriate for the convective boundary layer. We pick $N = 8$ and the chosen values of c_n are shown in Table A.2. The full velocity field (A.5) at an arbitrary point is obtained by interpolating the contribution from each flow. Finally, though our simulation constructs a cubic box of size $4z_i$, we constrain ourselves to the first quarter in the vertical direction for the symmetry reasons mentioned above. The resulting velocity profile and the Richardson superdiffusive law are shown in Fig. A.2.

A.2 Learning to soar: kinematic model

In the main text, we described our results for Rayleigh-Bénard flow. In this Section, we detail the results for the kinematic model described above. The gliders are trained using the same learning procedure and glider mechanics presented in the main text. The upshot is that learned policies are similar for the two flows and the main features of the flying policies discussed in the main text apply to the synthetic flow as well.

A.2.1 Setup

We consider a three-dimensional setting as described in the previous Section. The core mean flow (A.4) is centered at the origin with radius $R = 0.25$km and the maximal velocity at the center is set to 5m/s. Turbulent fluctuations of magnitude $u_{\text{rms}}^{\text{kin}}$ and the statistics described above are added on top of the Gaussian core. The fluctuations have long-range spatial correlations and the longest (and slowest) modes relax on timescales comparable to those of each ascent. The glider starts from the edge of the Gaussian thermal, facing away from its center, at an initial distance $\boldsymbol{r}_{\perp}^{\text{init}} = 2R$, and attempts to find the center of the thermal amidst turbulent fluctuations.

The groundspeed of a glider has three contributions :

$$\boldsymbol{u}^{\text{ground}} = \boldsymbol{v}^{\text{glider}} + \boldsymbol{u}^{\text{thermal}} + \boldsymbol{u}^{\text{kin}}, \tag{A.12}$$

where $\boldsymbol{v}^{\text{glider}}$ is the air velocity of the glider (see the Section on glider mechanics in the main text), $\boldsymbol{u}^{\text{thermal}}$ is the contribution due to the mean Gaussian core, and $\boldsymbol{u}^{\text{kin}}$ is the contribution due to the turbulent fluctuations. The airspeed and heading of the glider are controlled by the angle of attack and the bank angle of the glider (see main text). In this setting, we have three velocity scales: the contribution from the mean Gaussian core, the airspeed of the glider and the magnitude of fluctuations. Correspondingly, we have three regimes (I) the weak fluctuations regime, where the magnitude of fluctuations is smaller than the mean contribution of the core, (II) the strong fluctuations regime, where the mean is masked by fluctuations, and (III) the extreme fluctuations regime, when the fluctuations are larger than the airspeed of the glider. As a measure of the level of fluctuations, we define $\hat{u}_{\text{rms}} = u^{\text{kin}}_{\text{rms}}/u^{\text{thermal}}(\boldsymbol{r}_\perp = \boldsymbol{r}^{\text{init}}_\perp)$, i.e. the ratio between turbulent fluctuations and the thermal velocity at the starting point. In terms of $\hat{u}_{\text{rms}}$, the three regimes correspond to $\hat{u}_{\text{rms}} < 1$, $1 < \hat{u}_{\text{rms}} < 6$ and $\hat{u}_{\text{rms}} > 6$. We expect, as in the case of the RB flow, that the policy learned by the glider differs in the regimes of weak fluctuations and strong/extreme fluctuations. The turbulent fluctuations and glider flight are resolved with time steps of one second each, as for the results described in the main text. We used precisely the same architecture for the reinforcement learning algorithm as in the case of Rayleigh-Bénard flow (see main text).

A.2.2 Results

As for Rayleigh-Bénard flow, we found that the vertical wind acceleration $\boldsymbol{a}_z$ and the torques τ are the best mechanical cues to guide turbulent navigation. The angle of attack does not improve performance for reasons similar to those presented in the main text. We fix the set of observables to $\boldsymbol{a}_z$ and τ, and the glider has control over its bank angle. Figure A.3 shows the

learning curve and the average gain in height for different values of $\hat{u}_{\text{rms}}$. As anticipated, three regimes can be distinguished : (I) When $\hat{u}_{\text{rms}} < 1$, the contribution of the Gaussian core to the wind velocity is dominant, which makes its location and climbing relatively easy and allows for a large gain in height. (II) At $1 < \hat{u}_{\text{rms}} < 6$, fluctuations dominate and the glider is forced to learn how to exploit the fluctuations to soar. (III) When the fluctuations exceed the airspeed of the glider, the glider is carried away by the strong flow and progressively loses control over its trajectory. We observe indeed a declining gain in height for these extreme values of fluctuations.

Sample policies for $\hat{u}_{\text{rms}} = 0$ and $\hat{u}_{\text{rms}} = 5$ are shown in Figure A.4. Comparison with Figure 4 of the main text shows that the qualitative features of the policies in the two flows are very much the same; the arguments presented there directly apply to this case. The optimal bank angles can be obtained for each $(\boldsymbol{a}_z, \tau)$ pair by finding the mode of the distribution $\Pr(\mu_{t+1})|\boldsymbol{a}_z, \tau)$. Figure A.4 shows the optimal bank angles for the case of negative $\boldsymbol{a}_z$ and $\tau < 0$. The sharp change in policy occurs at the boundary between regimes II and III; note that these correspond to the weak and strong flow regimes for the Rayleigh-Bénard (RB) flow. The scaling of the boundary between "large" and "small" fluctuations ($-\hat{\boldsymbol{a}}_z$) is qualitatively similar but its nonlinear profile quantitatively differs from the linear one that we found for the RB flow. In Figure A.4D, we also show a simplified version of the heat map extended to a larger range. To obtain this figure, we proceed as in the maint text, i.e. defined a cutoff in the optimal bank angle at 12.5° that separates out the "large" and "small" fluctuation regions. The diverse colors shown in the figure correspond to simulations with expanded bin sizes in the tiled representation of $\boldsymbol{a}_z$.

A.3 Control over angle of attack during inter-thermal flight

As elucidated in the main text, for the task of finding and centering a single thermal, control over angle of attack offers minor improvement in the performance of the glider. However, angle of attack is expected to be relevant during inter-thermal flight, i.e. when the glider needs to

travel large distances quickly while avoiding the dangerous case of missing the thermal, losing height and crashing to the ground. The importance of angle of attack for the full case of inter-thermal flight in realistic turbulent conditions is a direction for future work (see the Discussion section of the main text). In this section, we considered a very simplified setting of inter-thermal flight and verified that control over angle of attack does indeed offer significant advantages.

We consider a glider constrained to moving in the X-Z plane, where Z is the vertical, for reasons that will be clear momentarily. The glider faces an X-dependent vertical velocity profile consisting of a downward current followed by a symmetric upward current as shown in panel A of Figure A.5. The net gain in height provided by the currents for a glider moving at a constant speed is zero. However, by modulating its angle of attack to slow down during regions of updraft and increasing its pace during regions of downdraft, the glider can achieve a net positive gain in height. The bank angle does not play any role here as the motion is constrained to the X-Z plane, which is the reason why we selected this geometry.

We used a reward function $\propto (\boldsymbol{u}_z - v_c)/v_\perp$, where $\boldsymbol{u}_z, v_c$ and $v_\perp$ are the vertical wind velocity, climb rate and horizontal speed respectively. The reward function encourages the glider to seek smaller horizontal speeds while ascending and larger horizontal speeds while sinking. The state space in this case includes the angle of attack discretized to 7 states between 2.5° and 17.5°, and the vertical wind velocity discretized to 25 bins between -12m/s and 12m/s. At each step, the glider has an option of increasing by 2.5°, decreasing by 2.5° or maintaining the same angle of attack. The bank angle is fixed at zero degrees. Panel B of Figure A.5 shows the net gain in height relative to a glider moving at fixed speed with the number of training episodes.

Even though the setting considered here is extremely simplified, results show the advantage provided by the control of the angle of attack. Quantifying the advantage and identifying the corresponding policy of control for realistic turbulent flows encountered during long-distance migration or cross-country glider competitions is the subject of ongoing work.

Table A.1: Values for the parameters employed in our simulations and training of the glider.

Label	Description	Value
Ra	Rayleigh number	10^8
Pr	Prandtl number	0.7
z_i	Inversion height	1km
mg/S	Wing loading	10N/m^2
η	Learning rate	0.1
γ	Discount factor (fixed)	0.98
τ_{temp}	Softmax "temperature" (early stages)	2.0
τ_{temp}	Softmax "temperature" (later stages)	0.2
l	Wingspan	10m
Δt	Time step	1s
v_{glider}	Glider airspeed (at fixed α)	4m/s
α	Angle of attack (fixed)	9°
$\boldsymbol{a}_z^{\text{thresh}}$	Threshold for vertical wind acceleration	0.05 m/s^2
τ^{thresh}	Threshold for torque	1 m^2/s

Table A.2: The parameters c_n (see Eq. (A.5)) used for the kinematic turbulence model.

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
1.1	0.8	0.7	0.7	0.6	0.6	0.5	2.0

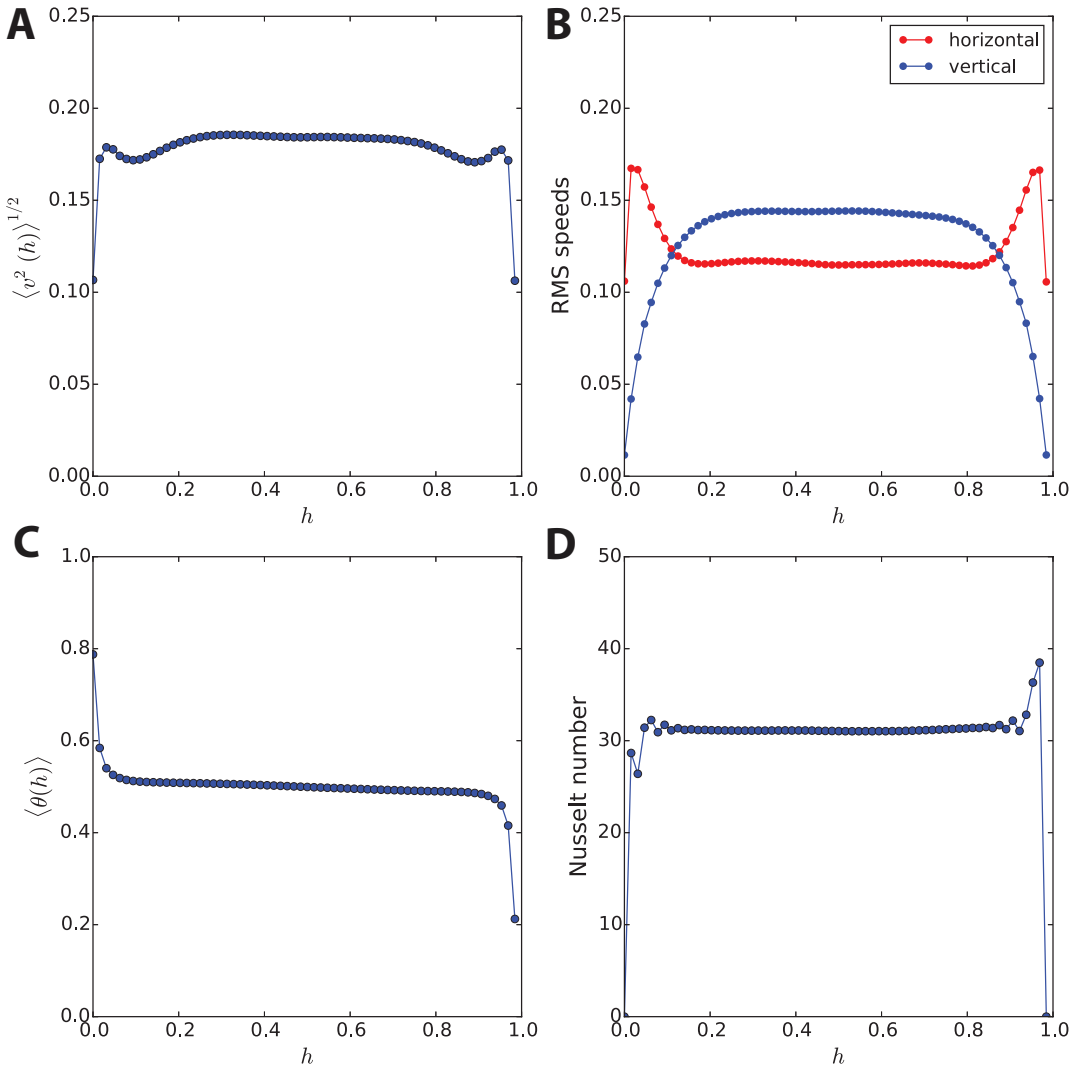

Figure A.1: Some additional observables for the Rayleigh-Bénard simulations. (A) the root-mean-square (rms) velocity ; (B) the horizontal and vertical rms velocities ; (C) the mean temperature ; (D) the profile of the Nusselt number *vs* height.

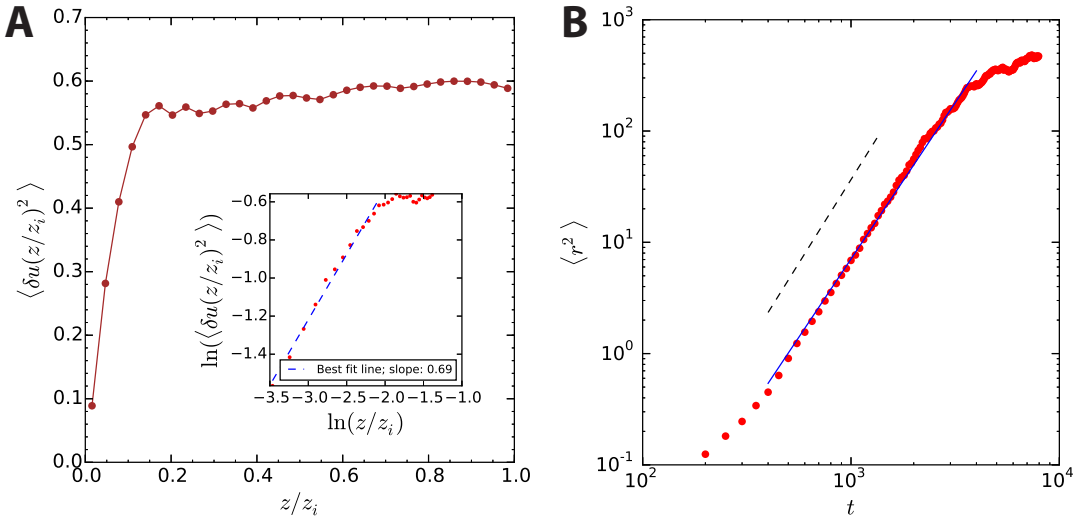

Figure A.2: Properties of the flow for the kinematic model of turbulence. (A) The mean-squared velocity profile. (B) The Richardson's superdiffusive law is well captured by our model. Small deviations are due to finite-size effects and the observed exponent is 2.7 (blue solid line).

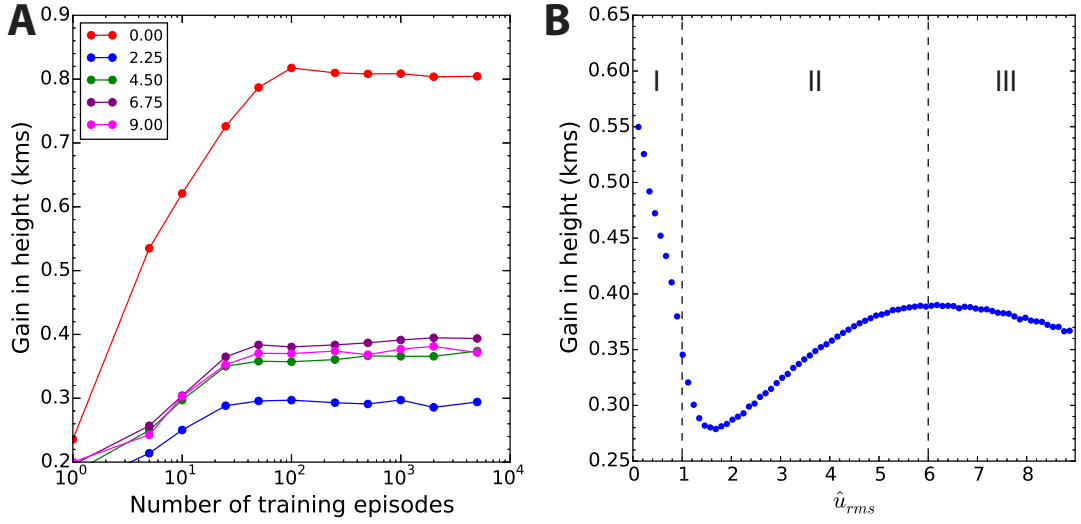

Figure A.3: Training and learning for the kinematic model of turbulence. (A) Learning curves for various values of $\hat{u}_{\text{rms}} = 0, 2.25, 4.5, 6.75, 9$. Vertical lines separate the three regimes of weak (I), strong (II) and extreme (III) fluctuations.

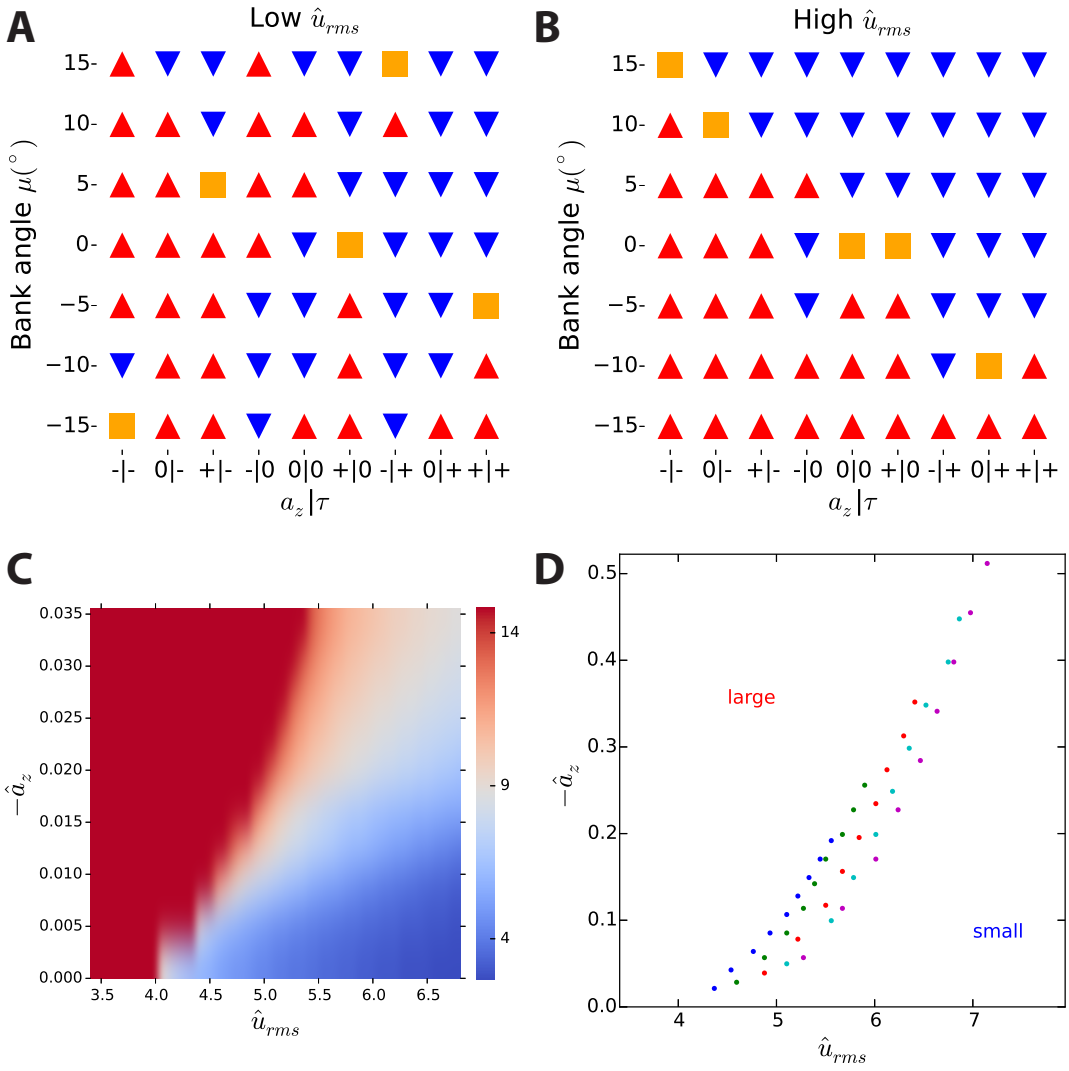

Figure A.4: Learned policies for the kinematic model of turbulence. Panels A and B show the learned policies at $\hat{u}_{\text{rms}} = 0$ and $\hat{u}_{\text{rms}} = 5$. Panel C shows a heat map of the optimal bank angles for negative $\boldsymbol{a}_z$ and $\tau < 0$. Panel D shows a simplified version of the heat map, similar to the main text.

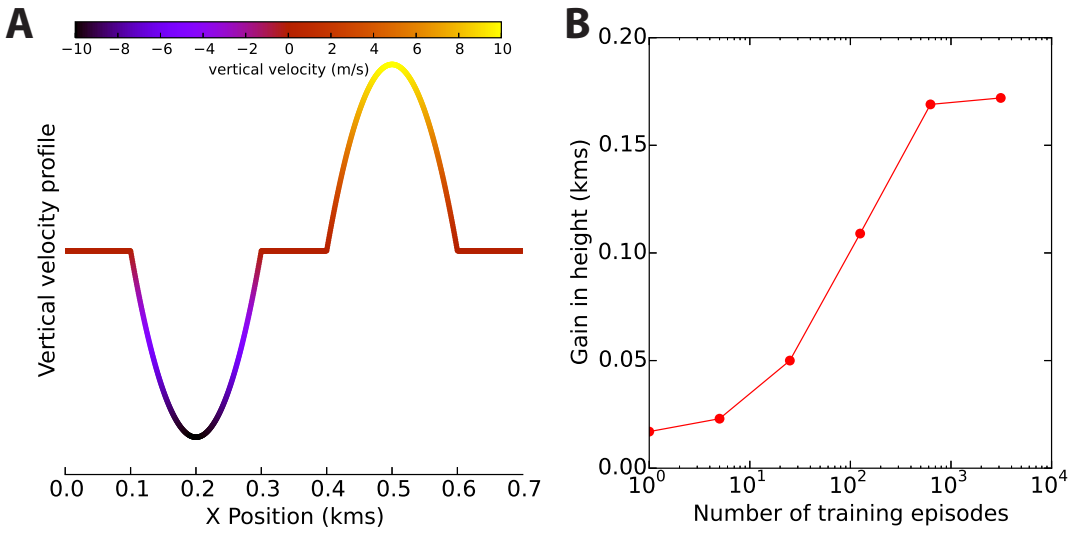

Figure A.5: Control over angle of attack during inter-thermal flight. Panel A shows the vertical wind velocity profile $\boldsymbol{u}_z(x)$ on the X-axis. Panel B shows the improvement in performance as training progresses, showing that the glider indeed learns to modulate its angle of attack for greater ascent.

Appendix B

Supplementary Information for Chapter 2

B.1 On-board estimation of the navigational cues

For a given desired pitch ϕ_d and desired bank angle μ_d, the flight controller implements a feedback control system such that:

$$\frac{d\phi}{dt} = \frac{\phi_d - \phi}{\tau}, \tag{B.1}$$

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau}, \tag{B.2}$$

where ϕ is the pitch (Figure B.2), μ is the bank angle and τ is a user-set time scale of control. ϕ_d is set fixed during flight and can be used to indirectly modulate the angle of attack, α, which determines the airspeed V and sink rate w.r.t air of the glider ($-v_z$). When α is large, the glider has a low airspeed and a low sink rate, while at small α the glider is faster but also sinks more rapidly. The glider's glide polar curve, relating the sink rate and airspeed for different values of α at equilibrium, depends on μ, the lift coefficient $C_L(\alpha)$ and the drag coefficient $C_D(\alpha)$ as

$$\frac{-v_z}{V} = \frac{C_D(\alpha)}{C_L(\alpha)\cos\mu}. \tag{B.3}$$

The ratio on the left hand side is called the glide angle γ (Figure B.2), where

$$\gamma \equiv -\boldsymbol{v}_z/V = \alpha - \phi. \tag{B.4}$$

The lift and drag coefficients are determined by the geometry of the plane; their form can be derived in certain simplified situations [23, 48]. Equations (B.3) and (B.4) together relate the measurable quantities, ϕ and the bank angle μ, to α at equilibrium. Actions of increasing, decreasing or keeping the same bank angle are taken in time steps of t_a by changing μ_d such that μ increases linearly from μ_i to μ_f in t_a:

$$\mu_d(t) = \mu_i + (\mu_f - \mu_i)\frac{t+\tau}{t_a}. \tag{B.5}$$

B.1.1 Estimation of the vertical wind acceleration

The vertical wind acceleration $\boldsymbol{a}_z$ is defined as:

$$\boldsymbol{a}_z \equiv \frac{d\boldsymbol{w}_z}{dt} = \frac{d}{dt}\left(\boldsymbol{u}_z - \boldsymbol{v}_z\right), \tag{B.6}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the velocities of the glider w.r.t the ground and air respectively, and $\boldsymbol{w}$ is the wind velocity. Here, we have used the relation

$$\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}. \tag{B.7}$$

An estimate of $\boldsymbol{u}$ is obtained in a straightforward manner from the EKF, which uses the GPS and barometer readings to form the estimate. On the other hand, the measurement of $\boldsymbol{v}_z$ is confounded by various aerodynamic effects that significantly affect it on relevant time scales of a few seconds. Artificial accelerations introduced due to these effects impair accurate estimation of the wind acceleration and thus alter the perceived state during decision-making and learning.

Two effects significantly influence variations in v_z: (1) Sustained pitch oscillations with a period of few seconds and a varying amplitude, and (2) Angle of attack variations, which occur in order to compensate for the imbalance of lift and weight while rolling. A full analysis of dynamic stability involves a set of four coupled differential equations [48] and is further complicated by the feedback control mechanism. We instead provide qualitative arguments and validate them using our data.

The longitudinal dynamic modes of the plane include short period oscillations and the phugoid. Short period oscillations are largely angle of attack variations, and the oscillations are usually heavily damped. Phugoid oscillations of longer period are less damped and are accompanied by oscillations of pitch at almost constant angle of attack. Using a reduced-order model of longitudinal stability [49], the time period of the phugoid oscillations can be estimated from the airspeed V as $\sqrt{2}\pi V/g \approx 3.5$s (here g is gravity and $V \approx 8$ m/s), which is consistent with the time period seen in experiments (Figure B.3A). Phugoidal oscillations are sustained due to constant perturbations to the pitch-wise moment when rolling. The amplitude and phase of the oscillations is determined by the magnitude and sign of $\dot{\mu}$ respectively. The amplitude is $\propto \mu^2$ and can be $> 5^\circ$ at bank angles of 30°. From (B.4), we see that pitch oscillations of a five degrees (~ 0.1 radian) at an airspeed $V = 8$m/s can give rise to a change in v_z of ~ 0.8m/s, which is of the same magnitude as the sink rate, and thus constitutes a significant contribution.

The lift-weight imbalance while rolling is compensated by a change in angle of attack. Suppose that a plane in equilibrium at bank angle μ_0, airspeed V_0 and angle of attack α_0 rolls to μ, V and α respectively. In equilibrium at μ_0 and μ, by balancing the forces along the vertical axis we get $L(\alpha_0)\cos\mu_0 = W = L(\alpha)\cos\mu$, where W is the weight of the glider. Here, the dependence of the lift on the angle of attack is emphasized (the contributions due to a non-zero glide angle are small and ignored here). Since $L(\alpha) \propto V^2 C_L(\alpha)$, this yields the relation $V^2 C_L(\alpha)\cos\mu = V_0^2 C_L(\alpha_0)\cos\mu_0$. Airspeed measurements in our experiments show that the change in V is negligible (Figure B.4A), and does not compensate for the change in lift. Instead,

the change in lift is largely balanced by a change in the angle of attack, so that:

$$\frac{C_L(\alpha)}{C_L(\alpha_0)} \approx \frac{\cos\mu_0}{\cos\mu}. \tag{B.8}$$

Below the stall angle, the lift coefficient is approximately a linear function $C_L(\alpha) = A(\alpha - \alpha_i)$, where α_i is usually negative and its value depends on the geometry and the angle of incidence of the wing. We thus obtain

$$\frac{\alpha - \alpha_i}{\alpha_0 - \alpha_i} \approx \frac{\cos\mu_0}{\cos\mu}, \tag{B.9}$$

$$\Delta\alpha \approx (\alpha_0 - \alpha_i)\left(\frac{\cos\mu_0}{\cos\mu} - 1\right), \tag{B.10}$$

where $\Delta\alpha \equiv \alpha - \alpha_0$.

Suppose a plane which is steady at zero bank angle has an angle of attack α_0, pitch ϕ_0 and vertical velocity w.r.t air of $\boldsymbol{v}_{z,0}$. The deviation of $\boldsymbol{v}_z$ from $\boldsymbol{v}_{z,0}$ for a particular bank angle at a given instant is (from (B.4))

$$\Delta\boldsymbol{v}_z = -\Delta V\,(\alpha_0 - \phi_0) - V\,(\Delta\alpha - \Delta\phi)\,. \tag{B.11}$$

Here $\Delta\alpha$ is assumed to depend on the instantaneous bank angle as given in equation (B.10), which is justified by our arguments that the longitudinal oscillations are phugoidal i.e., the angle of attack is not influenced. Since the change in V is small, the first term can be ignored and the second term can now be used as an approximation for the instantaneous $\boldsymbol{v}_z$ (up to a constant term) given the current bank angle and pitch, which are obtained from measurements. The constant term is ignored since our interest is in the derivative of $\boldsymbol{v}_z$.

In order to measure the variations in $\boldsymbol{v}_z$ in response to the glider's turn, we first observe that $\langle \boldsymbol{u}_z \rangle = \langle \boldsymbol{v}_z \rangle$ from (B.7) since $\langle \boldsymbol{w}_z \rangle = 0$. We compute $\langle \boldsymbol{u}_z \rangle$ (and thus $\langle \boldsymbol{v}_z \rangle$) by averaging the change in $\boldsymbol{u}_z$ (measured in the field) over hundreds of specific bank angle transitions. We verify

that changes in v_z over 13 possible bank angle transitions are indeed captured by (B.11) (Figure B.3B). Note that there is only one free parameter, $\alpha_0 - \alpha_i$, which is fit. The vertical wind velocity w_z is then estimated from (B.7).

The vertical wind acceleration a_z is smoothed using an exponential smoothing kernel with time scale σ_a. An exponential filter of time scale σ acts on an input x to give the smoothed output $\tilde{x}$ as,

$$\tilde{x}(t) = \int_{-\infty}^{t} x(s) e^{-\frac{t-s}{\sigma}} \sigma^{-1} ds, \tag{B.12}$$

where the tilde is hereafter used to denote quantities smoothed by an exponential filter. Substituting $a_z = \frac{dw_z}{dt}$ for x and integrating by parts, we get

$$\tilde{a}_z = \frac{w_z - \tilde{w}_z}{\sigma_a}, \tag{B.13}$$

In our implementation, another layer of exponential smoothing of smoothing time scale $\sigma'_a \ll \sigma_a$ is applied in order to average over sensory noise. As a consistency check, we verify that $\tilde{a}_z$ is unbiased for different bank angle transitions (Figure B.4B).

B.1.2 Estimation of vertical wind velocity gradients across the wings

Spatial gradients in the vertical wind velocity induce rolling moments on the plane, which we estimate using the deviation of the measured bank angle from the expected bank angle. The total rolling moment of the plane has contributions from three sources – (1) The feedback control of the plane, which acts according to equation (B.2), (2) The spatial gradients in the wind including turbulent fluctuations, and (3) Roll moments created due to various aerodynamic effects, which we detail below.

The latter two contributions create a dynamical effect that perturb the evolution of the

bank angle from equation (B.2). The modified evolution of the bank angle is modeled as

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau} + \omega(t) + \omega_{\text{aero}}(t), \tag{B.14}$$

where $\omega(t)$ and $\omega_{\text{aero}}(t)$ are contributions to the roll-wise angular velocity due to the wind and aerodynamic effects respectively. We empirically find four major contributions to $\omega_{\text{aero}}(t)$ - (1) the dihedral effect, which is a stabilizing moment due to the effects of sideslip on a dihedral wing geometry, (2) the overbanking effect, which is a destabilizing moment that occurs during turns with small radii, (3) Trim effects, which create a constant moment due to asymmetric lift on the two wings, and (4) a loss of rolling moment generated by the ailerons while rolling at low airspeeds.

Expanding the dihedral and overbanking effects around $\mu = 0$, their contributions to $\omega_{\text{aero}}(t)$ can be modeled with two terms of the form $-\mu/T_{\text{dih}}$ and $-\mu/T_{\text{ob}}$ respectively, with $T_{\text{dih}} > 0$ and $T_{\text{ob}} < 0$. The value of T_{dih} depends on the geometry of the wing and the airframe, whereas T_{ob} depends on the radius of the turns at μ. The radius of a turn at bank angle μ and airspeed V is given by

$$R = \frac{V^2}{g \tan \mu}, \tag{B.15}$$

For $V = 8$m/s and $\mu = 30^\circ$, the radius is ~ 10m. For wingspans of a few meters, typical of model sailplanes, the effect can be significant. The trim effect appears as a constant bias $-b$. The effective loss of rolling moment at low airspeeds is modeled as an additional term of the form $-\frac{\mu_d - \mu}{T_{\text{roll}}}$ that opposes changes in the bank angle towards the desired bank angle. In summary, an unbiased estimation of the torque requires a calibration of three parameters related to the aerodynamics of the glider - $T_s^{-1} \equiv T_{\text{dih}}^{-1} + T_{\text{ob}}^{-1}$, T_{roll} and b. The full equation for the evolution of

the bank angle is now written as:

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau} - \frac{\mu}{T_s} - \frac{\mu_d - \mu}{T_{\text{roll}}} - b + \omega(t), \tag{B.16}$$

The three parameters are measured by making repeated transitions between bank angles of $0^\circ, \pm 15^\circ, \pm 30^\circ$ by increasing, decreasing the bank angle by 15° or keeping same angle over intervals of 3 seconds. Averaging the bank angle in this interval over many such transitions yields the evolution of the bank angle without the wind contribution in (B.16) (Figure B.5B). The averaged (B.16) can be integrated exactly to get an analytical form for the bank angle. For linear transitions from μ_i to μ_f in a time interval t_a, plugging (B.5) into the averaged (B.16) and integrating leads to

$$\begin{aligned} \mu(t) = \frac{\tau' t \Delta\mu}{\tau'' t_a} + \left(\tau''^{-1}\mu_i - b\right)\tau' e^{-t/\tau'} + \\ + \left(\frac{\tau\Delta\mu}{\tau'' t_a} - b - \frac{\tau'\Delta\mu}{\tau'' t_a}\right)\tau'\left(1 - e^{-t/\tau'}\right), \end{aligned} \tag{B.17}$$

where we have defined $\tau'^{-1} = \tau^{-1} - T_{\text{roll}}^{-1} + T_c^{-1}$, $\tau''^{-1} = \tau'^{-1} - T_c^{-1}$ and $\Delta\mu = \mu_f - \mu_i$. The three parameters are then obtained by fitting the predicted curves from the above equation to the 13 experimentally obtained bank angle transition curves, as shown in Figure B.5A.

The roll-wise torque is smoothed over a time scale σ_ω using (B.12) to obtain the equation for the smoothed torque $\tilde{\omega}$:

$$\tilde{\omega}(t) = \frac{\mu - \tilde{\mu}}{\sigma_\omega} + \frac{\tilde{\mu}}{T_s} - (\tilde{\mu}_d - \tilde{\mu})\left(\frac{1}{\tau} - \frac{1}{T_{\text{roll}}}\right) + b, \tag{B.18}$$

where we again use the tilde to denote quantities smoothed over the time scale σ_ω using (B.12). As in the case of a_z, another layer of smoothing of time scale σ'_w is applied. We find that the bias b can change across different flights of the same glider. The bias is estimated on-board before the soaring algorithm is activated by exponentially averaging the torque uncorrected for bias

over a time scale of two minutes. Finally, we verify that the estimated ω for different bank angle transitions is indeed unbiased, as shown in Figure B.5B.

B.2 Reward shaping and policy invariance

Reinforcement learning algorithms [12] are typically posed in the framework of a Markov Decision Process (MDP). In an MDP, an agent traverses a state space by taking actions while receiving associated rewards. A transition matrix, denoted by $T(s'|s,a)$, gives the probability of transitioning to a particular state s' given the agent's current state s and its current action a and encodes the statistics of the environment and its interactions with the agent. The reward function $R(s,a)$ defines the expected reward given when action a is taken in state s. The agent's control over actions is represented by its policy $\pi(a|s)$, which is the probability that the agent takes action a at state s. The expected discounted sum of future rewards for a particular state-action pair (s,a) is given by the Q function, which is written here in a recursive form:

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s',a'} T(s'|s,a)\pi(a'|s')Q_\pi(s',a'). \tag{B.19}$$

Here $\gamma\,(0 \leq \gamma < 1)$ is the discount factor, which determines the time scale of future rewards the agent cares about, and the subscript is used to highlight that the Q values depend on the policy π.

To train the glider, we choose the local vertical wind acceleration $\boldsymbol{a}_z$ as our reward function. The choice of $\boldsymbol{a}_z$ as an appropriate reward signal is motivated by observations made in simulations from [42]. In general, multiple reward functions can lead to the same policy, which opens the possibility for *reward shaping*, where a reward function modified from that of the underlying MDP is chosen in order to accelerate the learning process [46]. Reward shaping is particularly useful when the reward is delayed and learning is encumbered by the credit assignment problem. For the purpose of soaring, we aim to maximize the expected gain in height over a time interval of a few minutes. An intuitive choice for the reward function would then be the local vertical wind

velocity $\boldsymbol{w}_z$, in which case the the RL algorithm maximizes the quantity $\langle \sum_{i=0}^{\infty} \boldsymbol{w}_z(t_i)\gamma^i \rangle$, where t_i is the time of the ith time step and the angular brackets denote expectation values. In the limit of $\gamma \to 1$, this quantity is the expected gain in height over a time interval $\sim (1-\gamma)^{-1}$. However, we find that the choice of $\boldsymbol{w}_z$ as the reward function fails to drive learning in the soaring problem, possibly because the velocities are strongly correlated across states and their temporal statistics fails to satisfy the Markovian assumption. To justify our choice of $\boldsymbol{a}_z$ as the reward function, we show here that a policy π that is optimal for an MDP with expected reward $R(s,a)$ is also optimal for the same MDP with reward $\propto \langle R(s',a') \rangle_{s,a,\pi} - \gamma R(s,a)$, where $\langle R(s',a') \rangle_{s,a,\pi}$ is the expected reward at the next time step given by

$$\langle R(s',a') \rangle_{s,a,\pi} = \sum_{s',a'} R(s',a') T(s'|s,a) \pi(a'|s'). \tag{B.20}$$

Intuitively, this implies that using a particular reward function for an MDP is equivalent to using any "derivative" of the reward function as its proxy, where the derivatives are defined in the discounted difference sense as above. We first observe that the policy induces a Markov chain on the MDP defined by the transition probabilities

$$T_\pi(s'|s) = \sum_a T(s'|s,a) \pi(a|s). \tag{B.21}$$

The key assumption we make here is that the induced Markov chain has a stationary distribution ρ_π given by

$$\rho_\pi(s') = \sum_s T_\pi(s'|s) \rho_\pi(s) \tag{B.22}$$

The expected sum of future rewards for policy π is then

$$\begin{aligned}\mathcal{V}_\pi &= \sum_{s,a} \rho_\pi(s)\pi(a|s)Q_\pi(s,a), \\ &= \sum_{s,a} \rho_\pi(s)\pi(a|s)\left(R(s,a)+\gamma\sum_{s',a'} T(s'|s,a)\pi(a'|s')Q_\pi(s',a')\right), \\ &= \mathcal{R}_\pi + \gamma\sum_{s',a'} \rho_\pi(s')\pi(a'|s')Q_\pi(s',a'), \\ &= \mathcal{R}_\pi + \gamma\mathcal{V}_\pi,\end{aligned} \tag{B.23}$$

where we have defined the expected immediate reward, $\mathcal{R}_\pi = \sum_{s,a} \rho_\pi(s)\pi(a|s)R(s,a)$. The second step above follows from (B.19) whereas the third step uses the relation (B.22). We then have

$$\mathcal{R}_\pi = (1-\gamma)\,\mathcal{V}_\pi. \tag{B.24}$$

We wish to show that the expected sum of future rewards $\tilde{\mathcal{V}}_\pi$ for the MDP with reward function $\langle R(s',a')\rangle_{s,a,\pi} - \gamma R(s,a)$ is directly related to $\mathcal{V}_\pi$. The expected immediate reward $\tilde{\mathcal{R}}_\pi$ for this new process is given by (from (B.20))

$$\begin{aligned}\tilde{\mathcal{R}}_\pi &= \sum_{s,a} \rho_\pi(s)\pi(a|s)\left(\sum_{s',a'} T(s'|s,a)\pi(a'|s')R(s',a') - \gamma R(s,a)\right), \\ &= (1-\gamma)\mathcal{R}_\pi,\end{aligned} \tag{B.25}$$

where the second step is derived in a fashion similar to the third and fourth steps in (B.23). Since relation (B.24) also holds for the new MDP, we have $\tilde{\mathcal{R}}_\pi = (1-\gamma)\,\tilde{\mathcal{V}}_\pi$, which yields (together with (B.24) and (B.25)) that

$$\tilde{\mathcal{V}}_\pi = (1-\gamma)\mathcal{V}_\pi. \tag{B.26}$$

The above relation holds for *any* policy. In particular, it holds for the optimal policy π^*, which maximizes $\tilde{\mathcal{V}}_\pi$, and therefore is also the policy that maximizes $\mathcal{V}_\pi$.

B.3 Noisy gradient sensing in the turbulent atmospheric boundary layer

The navigational cues $\boldsymbol{a}_z$ and ω measure the gradients in the vertical wind velocity along and perpendicular to the heading of the glider. Updrafts and downdrafts are relatively stable structures in a varying turbulent environment. Thermal detection through gradient sensing constitutes a discrimination problem of deciding whether a thermal is present or absent given recent $\boldsymbol{a}_z$ and ω values. In this section, we estimate the magnitude of 'noise' due to turbulence that unavoidably accompanies gradient sensing in the atmospheric boundary layer. Similar estimates of the noise due to the statistical properties of the surrounding physical environment have been made for the sensing of concentration gradients by motile bacteria finding nutrients via chemotaxis [50]. There, the noise arises due to the properties of diffusion; slow diffusion of the few ligands present in the local environment of the cell leads to repeated binding of the same ligands on the cell's receptors, resulting in a biased estimate of the local mean concentration. We consider, as in [50], the case of a 'perfect instrument', which perfectly measures the local vertical wind velocity at its location and across its wings with no accompanying measurement noise and aerodynamics-induced bias.

In the Methods section, we estimate the SNR using simple scaling arguments. Here, we validate our scaling arguments by explicitly computing the SNR for the case of ω estimation. The calculation for $\boldsymbol{a}_z$ estimation is similar and we omit it for the sake of conciseness. Note that the estimate below is still accurate only up to constant factors of order unity.

The instantaneous rate of rotation or 'torque' generated by the vertical component of the fluctuating velocity field $\boldsymbol{w}(\boldsymbol{r},t)$ is given by $\left(\boldsymbol{w}_z^+ - \boldsymbol{w}_z^-\right)/l$, where the $\pm$ superscripts denote

the vertical wind velocities on the two wings and we have $\langle \boldsymbol{w}_z^{\pm} \rangle = 0$. We assume the torque is averaged over a time scale T using an exponential kernel, as in (B.12). We expect the dependence of the noise on l,V and T to remain invariant to the specific computation performed in integrating the torque; using an exponential kernel is convenient for simplifying the calculations. Suppose the glider moves with a fixed velocity $\boldsymbol{V}$ and the unit vector along the wings is $\hat{\boldsymbol{y}}$ (note that $\boldsymbol{V}$ and $\hat{\boldsymbol{y}}$ are perpendicular to each other). For ease of notation, suppose also that at the final time t the glider is at the origin. We have

$$l\tilde{\omega}(t) = \tilde{\boldsymbol{w}}_z^{+}(t) - \tilde{\boldsymbol{w}}_z^{-}(t), \tag{B.27}$$

where

$$\tilde{\boldsymbol{w}}_z^{\pm}(t) = \frac{1}{T}\int_{-\infty}^{t} \boldsymbol{w}_z\left(-\boldsymbol{V}(t-s) \pm \frac{l}{2}\hat{\boldsymbol{y}}, s\right) e^{-\frac{t-s}{T}} ds. \tag{B.28}$$

The variance $\delta\tilde{\omega}^2 = \langle \tilde{\omega}^2 \rangle$ is

$$l^2\delta\tilde{\omega}^2 = \langle \tilde{\boldsymbol{w}}_z^{+2} \rangle + \langle \tilde{\boldsymbol{w}}_z^{-2} \rangle - 2\langle \tilde{\boldsymbol{w}}_z^{+} \tilde{\boldsymbol{w}}_z^{-} \rangle. \tag{B.29}$$

We have $\langle \tilde{\boldsymbol{w}}_z^{+2} \rangle = \langle \tilde{\boldsymbol{w}}_z^{-2} \rangle$, where

$$\langle \tilde{\boldsymbol{w}}_z^{+2} \rangle = \frac{1}{T^2}\int_{-\infty}^{t}\int_{-\infty}^{t} \left\langle \boldsymbol{w}_z\left(-\boldsymbol{V}(t-s) + \frac{l}{2}\hat{\boldsymbol{y}}, s\right) \boldsymbol{w}_z\left(-\boldsymbol{V}(t-s') + \frac{l}{2}\hat{\boldsymbol{y}}, s'\right) \right\rangle e^{-\frac{t-(s+s')/2}{T/2}} ds ds', \tag{B.30}$$

$$= \frac{w^2}{T^2}\int_{-\infty}^{t}\int_{-\infty}^{t} \left(1 - \left(\frac{V|s-s'|}{L}\right)^{2/3}\right) e^{-\frac{t-(s+s')/2}{T/2}} ds ds', \tag{B.31}$$

where L is the length scale of the ABL and w is the magnitude of vertical wind velocity fluctuations $\langle \boldsymbol{w}_z^2 \rangle^{1/2}$. Here, we have used the two-point velocity correlation function in the turbulent regime, $\langle (\boldsymbol{w}(\boldsymbol{r}) - \boldsymbol{w}(\boldsymbol{r}'))^2 \rangle \sim |\boldsymbol{r} - \boldsymbol{r}'|^{2/3}$. Further, for V much larger than the velocity scale of the eddies

w, any decorrelation of wind velocities is due to the glider's motion. We can then assume that the eddies are frozen in time, which allows us to approximate the spatio-temporal correlations in the equations above using only the spatial component of the two-point correlation function. The above integral is simplified by transforming variables to $p = (s+s')/2, q = (s-s')/2$:

$$\langle \tilde{\boldsymbol{w}}_z^{+2} \rangle = \frac{2w^2}{T^2} \int_{-\infty}^{\infty} \left(1 - \left(\frac{2V|q|}{L}\right)^{2/3}\right) dq \int_{-\infty}^{t-|q|} e^{-\frac{t-p}{T/2}} dp \tag{B.32}$$

$$= \frac{2w^2}{T} \int_0^{\infty} \left(1 - \left(\frac{2Vq}{L}\right)^{2/3}\right) e^{-\frac{2q}{T}} dq \tag{B.33}$$

$$= w^2 \left(1 - \Gamma(5/3) \left(\frac{VT}{L}\right)^{2/3}\right) \tag{B.34}$$

The calculation of the last term in the RHS of (B.29) follows in a similar manner. We have

$$\langle \tilde{\boldsymbol{w}}_z^{+} \tilde{\boldsymbol{w}}_z^{-} \rangle = \frac{w^2}{T^2} \int_{-\infty}^{t} \int_{-\infty}^{t} \left(1 - \left(\frac{\left(l^2 + V^2|s-s'|^2\right)^{1/2}}{L}\right)^{2/3}\right) e^{-\frac{t-(s+s')/2}{T/2}} ds ds'. \tag{B.35}$$

Using the same transformation as above, and performing a straightforward calculation with $q' = 2q/T$, we get

$$\langle \tilde{\boldsymbol{w}}_z^{+} \tilde{\boldsymbol{w}}_z^{-} \rangle = w^2 \left(1 - \left(\frac{VT}{L}\right)^{2/3} \int_0^{\infty} \left(\alpha^2 + q'^2\right)^{1/3} e^{-q'} dq'\right) \tag{B.36}$$

where we have substituted $\alpha = l/VT$. Combining (B.29), (B.32) and (B.36), we get

$$l^2 \delta\tilde{\omega}^2 = 2w^2 \left(\frac{VT}{L}\right)^{2/3} \left(\int_0^{\infty} \left(\alpha^2 + q'^2\right)^{1/3} e^{-q'} dq' - \Gamma(5/3)\right) \tag{B.37}$$

The integral above can be found in [51] and is expressed in terms of the Struve functions, $\boldsymbol{H}_{\nu}$,

and the Bessel functions of 2nd kind, N_ν, to get

$$l^2\delta\tilde{\omega}^2 = 2w^2\left(\frac{VT}{L}\right)^{2/3}\left(\alpha^{5/6}2^{-1/6}\Gamma(1/2)\Gamma(4/3)\left(\boldsymbol{H}_{5/6}(\alpha)-N_{5/6}(\alpha)\right)-\Gamma(5/3)\right) \quad \text{(B.38)}$$

For $\alpha \ll 1$, the first terms of the series expansions of $\boldsymbol{H}_\nu$ and N_ν can be used to verify the scaling obtained from the above arguments. It is convenient to express N_ν in terms of the Bessel functions of the first kind, J_ν, as $N_\nu(x) = \left(\cos(\nu\pi)J_\nu(x) - J_{-\nu}(x)\right)/\sin(\nu\pi)$ and expand $J_{\pm\nu}(x)$ for small x. After a straightforward but lengthy calculation involving Gamma function identities we arrive for $\alpha \ll 1$:

$$l^2\delta\tilde{\omega}^2 = w^2\left(\frac{VT}{L}\right)^{2/3}\alpha^{5/3}\frac{\sqrt{3}\Gamma(1/3)\Gamma(1/6)}{\Gamma(1/2)} \quad \text{(B.39)}$$

The mean vertical velocity difference for a glider travelling tangential to a thermal having profile $\boldsymbol{W}_z$ is $|l\hat{\boldsymbol{y}}.\frac{\partial \boldsymbol{W}_z}{\partial \boldsymbol{r}}| \sim lW/R$ where W is the strength of the thermal and R its size. The signal to noise ratio for $\alpha \ll 1$ is therefore

$$\frac{|l\hat{\boldsymbol{y}}.\partial \boldsymbol{W}_z/\partial \boldsymbol{r}|}{l\delta\tilde{\omega}} \sim \frac{WV^{1/2}T^{1/2}l^{1/6}L^{1/3}}{wR}. \quad \text{(B.40)}$$

Plugging in typical values: W = 2 m/s, R = 50m, w = 0.5 m/s, l = 2m, V = 8 m/s, T = 3s, L = 1 km, we obtain an SNR of ~ 4. A similar calculation can be performed for the accelerations. For a glider moving towards a thermal as above, using the arguments above and simple dimensional considerations, we have

$$\frac{|\boldsymbol{V}.\partial \boldsymbol{W}_z/\partial \boldsymbol{r}|}{\delta\tilde{\boldsymbol{a}}_z} \sim \frac{WV^{2/3}T^{2/3}L^{1/3}}{wR}. \quad \text{(B.41)}$$

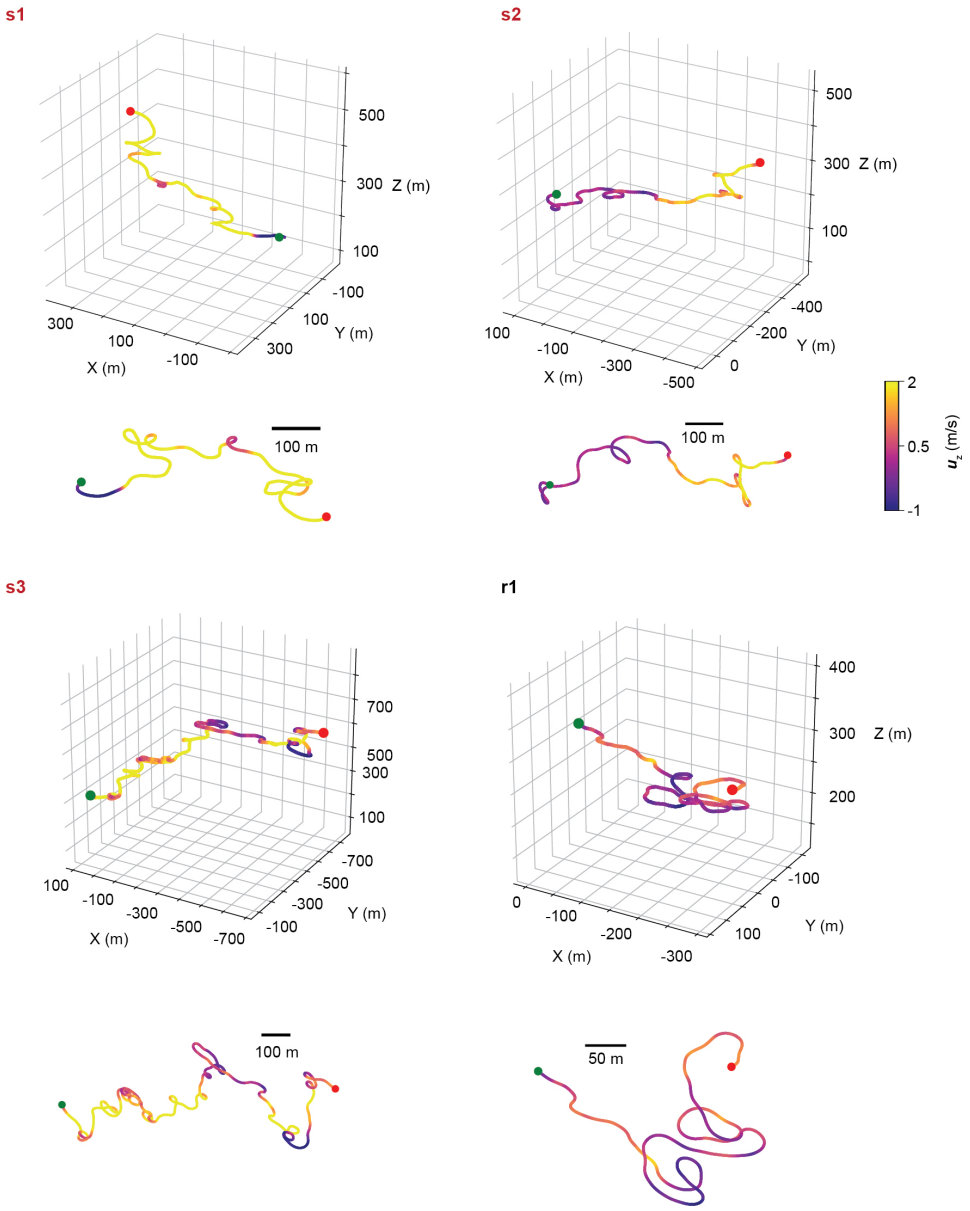

Figure B.1: Sample trajectories obtained in the field (3D and top view) with a glider using the learned thermalling strategy (labeled S) or a random policy that takes actions with equal probability (labeled R). The green (red) dot shows the start (end) point of the trajectory.

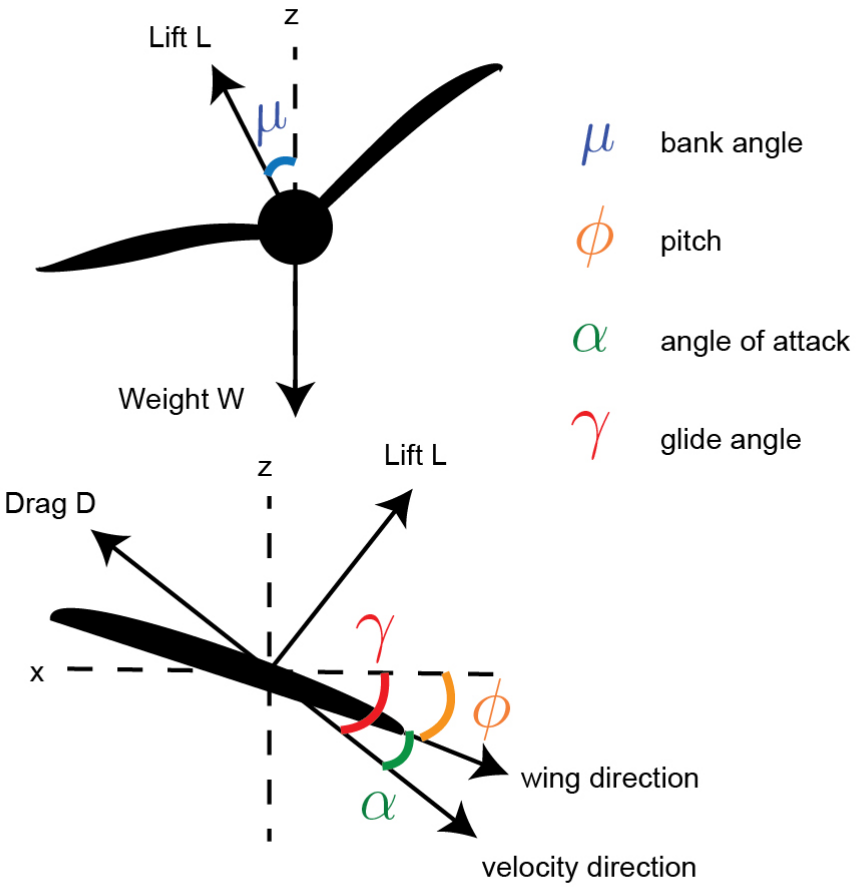

Figure B.2: The forces on a glider and the definitions of the various angles that determine the glider's motion.

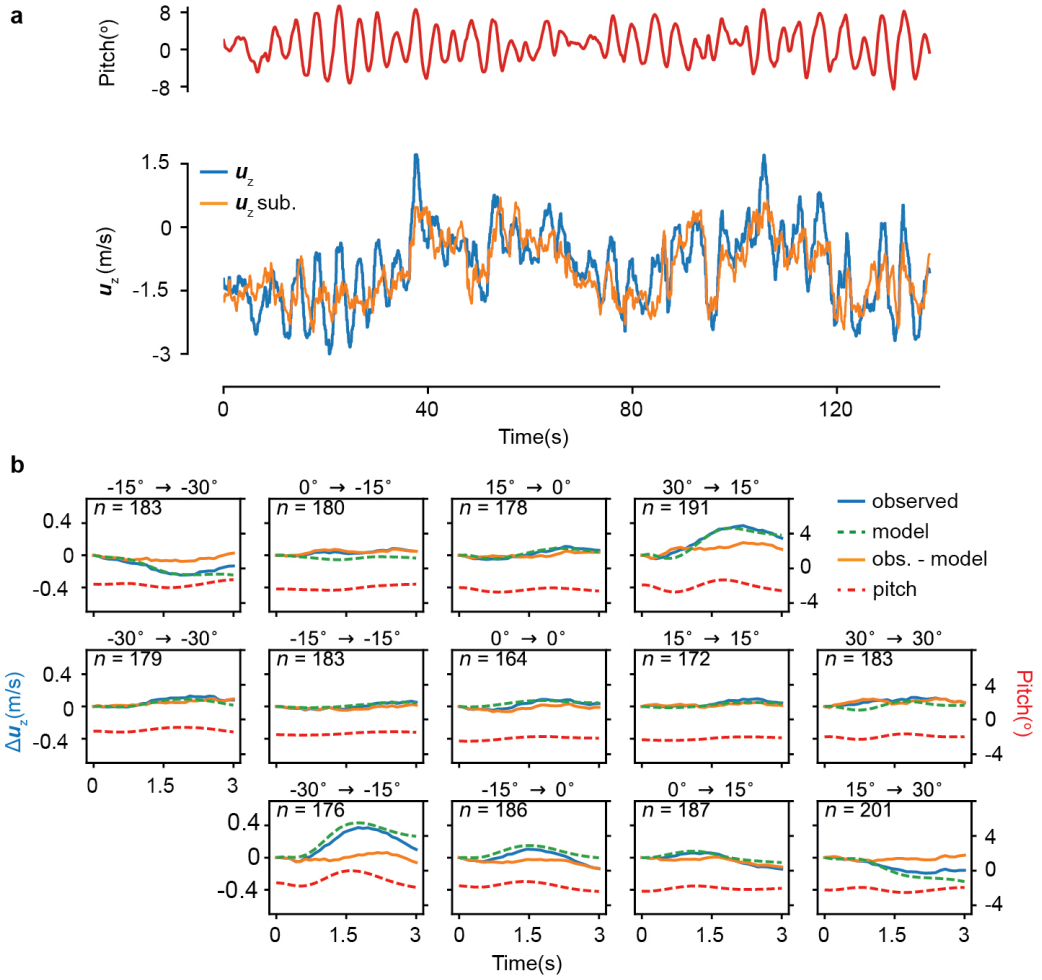

Figure B.3: (a) A trajectory of a glider's pitch and u_z. (b) The blue line shows the average change in u_z for each action. The green, dashed line shows the prediction from the model and the orange line is the estimated w_z. The right axis shows the averaged pitch as a red, dashed line.

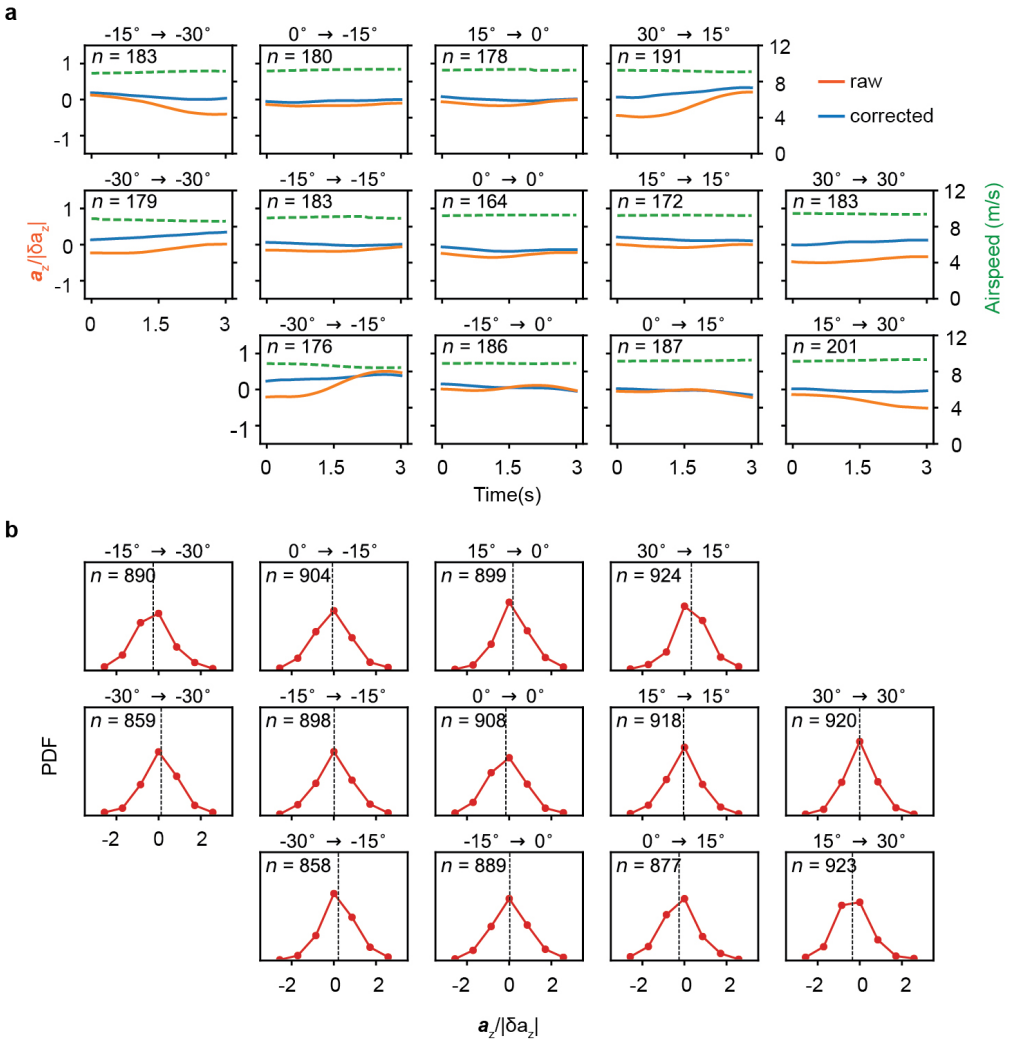

Figure B.4: (a) a_z plotted as in Figure B.3B, is shown in orange with (blue line) and without (orange line). The axis on the right shows the airspeed as a green, dashed line. (b) The PDFs of a_z for the different bank angle changes. The black, dashed line shows the median.

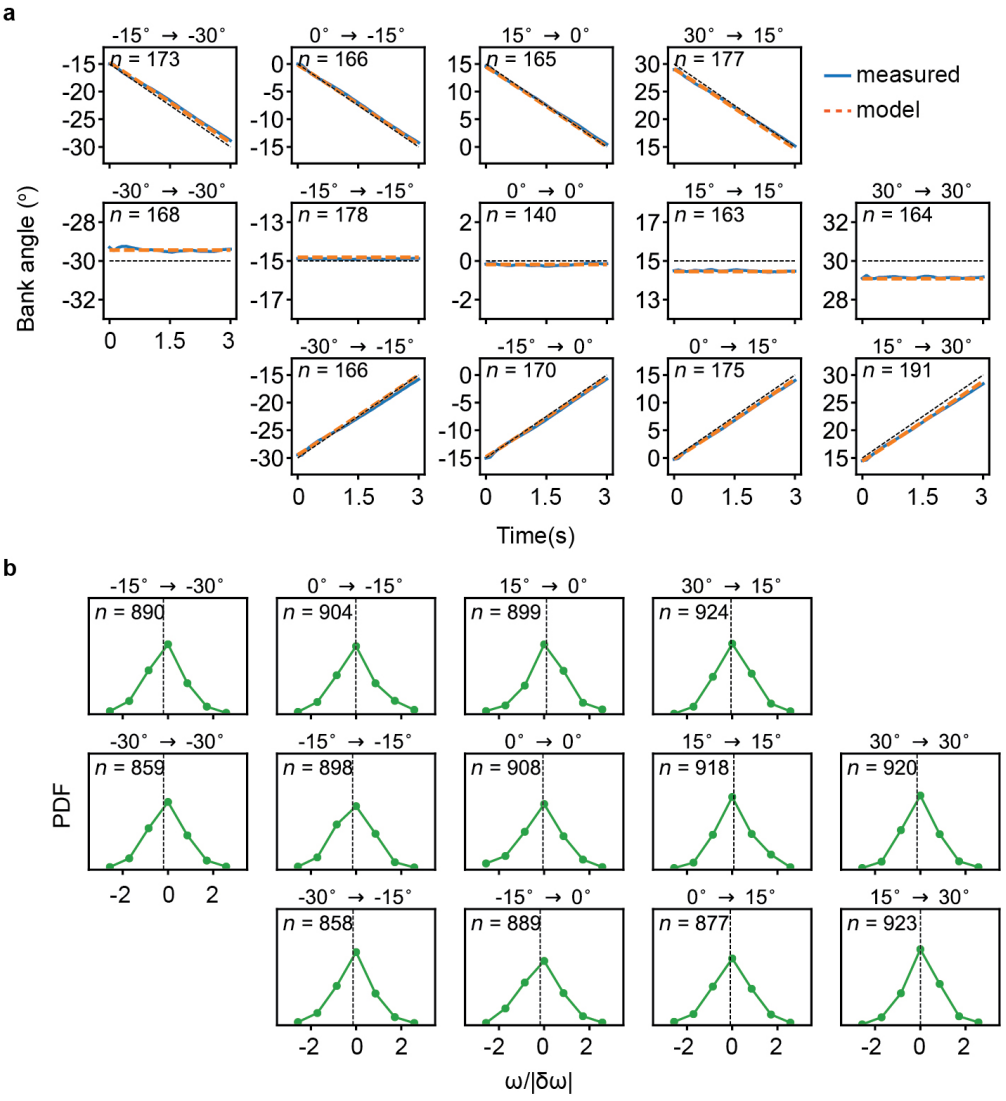

Figure B.5: (a) The averaged evolution of the bank angle shown as in Figure B.3B. The blue line shows the measured bank angle and the dashed, orange line shows the best fit line. (b) The PDFs of the torque ω for the different bank angle changes. The black, dashed line shows the median value.

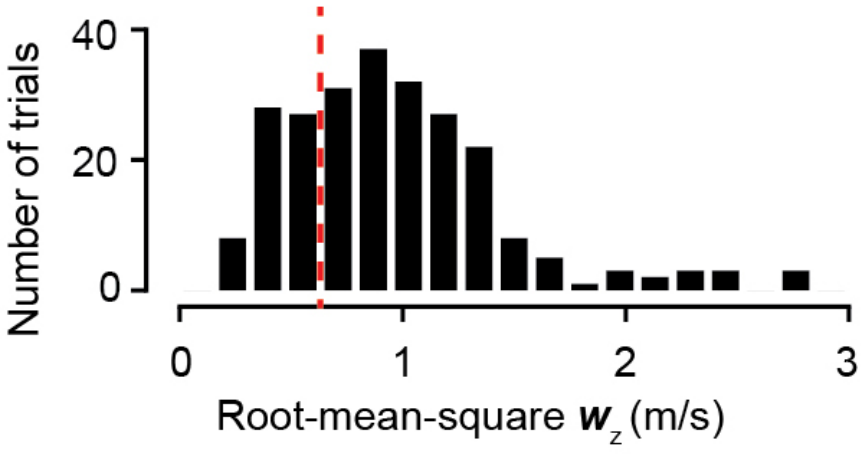

Figure B.6: The distribution of the strength of vertical currents observed in the field. The data is pooled from ∼240 3-minute trials collected over 9 days. The dashed, red line shows the threshold criterion imposed when measuring the performance of the strategy in the field (see Methods).

Table B.1: The parameter values used in the experiments performed in the field.

Label	Description	Value
l	Wingspan of glider	2m
ϕ_d	Desired pitch	-2°
τ	Feedback control time scale	0.45s
t_a	Interval between actions (learning)	3s
t_a	Interval between actions (soaring)	1.5s
$\alpha_0 - \alpha_i$	Net angle of attack	14°
V	Airspeed (typical)	6 to 8 m/s
T_{dih}	Dihedral effect timescale (typical)	14 to 30 s
T_{ob}	Overbanking effect timescale (typical)	-20 to -∞s
b	Trim bias (typical)	-2 to +2°/s
T_{roll}	Opposing roll timescale (typical)	1.5 to 3 s
$\pm K_a, \pm K_w$	Thresholds for $\boldsymbol{a}_z$ and ω state estimation	0.8 $\times$ std. dev
σ_a, σ_a	Exponential smoothing timescales for $\boldsymbol{a}_z$	$8t_a/3, 2ta/3$
σ_w, σ'_w	Exponential smoothing timescales for ω	$t_a, t_a/4$
γ	Discount factor for RL implementation	0.8

Bibliography

[1] Clarence D. Cone Jr. (1962) Thermal soaring of birds, *American Scientist*, Vol. 50, No. 1, pp. 180-209.

[2] Pennycuick, C. J. (1983) Thermal soaring compared in three dissimilar tropical bird species, Fregata magnificens, Pelecanus occidentalis and Coragyps atratus. *J Exp Biol* 102:307-325.

[3] Ehrlich P., Dobkin D. & Wheye D., *The Birder's Handbook: A Field Guide to the Natural History of North American Birds*, Touchstone, 1988.

[4] Newton I., *The Migration Ecology of Birds*, Academic Press, 2007

[5] Allen, M. J. (2007) Guidance and Control of an Autonomous Soaring UAV, *AIAA*, 2007-867.

[6] Garrat, J. R. *The Atmospheric Boundary Layer*, Cambridge Atmospheric and Space Science Series, 1994.

[7] Lenschow, D.H. & Stephens, P.L. (1980). The role of thermals in the convective boundary layer, *Boundary-Layer Meteorology*, 19(4), pp.509-532

[8] Young, G.S., (1988) Convection in the atmospheric boundary layer, *Earth-Science Reviews*, 25(3), pp.179-198.

[9] Ahlers G., Grossmann S., & Lohse D., (2009) Heat transfer and large scale dynamics in turbulent Rayleigh-Benard convection, *Rev. Mod. Phys.*, 81, 503.

[10] Frisch, U. *Turbulence: The Legacy of A. N. Kolmogorov*, Cambridge University Press, 1995.

[11] Akos, Z., Nagy, M. & Vicsek T. (2008) Comparing bird and human soaring strategies, *Proc. Natl. Acad. Sci. USA*, Vol. 105, No. 11:4139-4143.

[12] Sutton, R. S. & Barto, A. G. *Reinforcement learning : an introduction*, MIT Press, 1998.

[13] Wharington, J. & Herszberg, I. (1998) Control of a High Endurance Unmanned Aerial Vehicle, *ICAS*, 98-3,7,1.

[14] Reichmann, H. (1988) Cross-Country Soaring, *Thomson Publications*, Santa Monica, CA

[15] Doncieux, S., Mouret, J. B., & Meyer, J-A. (2007) Soaring behaviors in UAVs : 'animat' design methodology and current results. In *7th European Micro Air Vehicle Conference (MAV07)*, Toulouse

[16] Woodbury, T., Dunn, C. & Valasek, J. (2014) Autonomous Soaring Using Reinforcement Learning for Trajectory Generation, *AIAA* 2014-0990

[17] Akos, Z., Nagy, M., Leven, S. & Vicsek, T. (2010) Thermal soaring flight of birds and UAVs, *Bioinspir. Biomim.*, 5 045003.

[18] Popinet S. (2003) Gerris: A Tree-Based Adaptive Solver For The Incompressible Euler Equations In Complex Geometries *J. Comp. Phys*, 190, 572-600.

[19] Verzicco R. & Camussi R., (2003) Numerical experiments on strongly turbulent thermal convection in a slender cylindrical cell *J. Fluid. Mech*. Vol. 477, pp 19-49.

[20] Verzicco R. & Camussi R., (1999) Prandtl number effects in convective turbulence *J. Fluid. Mech*. Vol. 383, pp 55-73

[21] Fung, J. C. H., Hunt, J. C. R., Malik, N. A. & Perkins, R. J., (1992), Kinematic simulation of homogeneous turbulence by unsteady random Fourier modes, *J. Fluid Mech.*, vol. 236, pp. 281-318.

[22] von Mises R. *Theory of Flight*, McGraw Hill, 1945.

[23] Anderson J.R. Jr. *Introduction to Flight*, McGraw Hill, 1978.

[24] von Karman T., *Aerodynamics*, McGraw-Hill, 1963

[25] Tesauro, G. (1995) Temporal Difference Learning and TD-Gammon, *Commun. ACM*, 38, 58-68.

[26] Montague, P. R., Dayan, P. & Sejnowski, T. J. (1996) A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning, *Journal of Neuroscience*, 16(5), 1936-1947.

[27] Shraiman BI & Siggia ED, (2000), Scalar turbulence, *Nature*, 405, 639-646.

[28] Falkovich G., Gawedzki K. & Vergassola M, (2001) Particles and fields in fluid turbulence *Rev. Mod. Phys.*, 73, 913-975.

[29] MacCready, P. B. J. (1958) Optimum airspeed selector, *Soaring*, 10-11.

[30] Horvitz, N., Sapir N., Liechti F., Avissar R., Mahrer I., & Ran Nathan. (2014) The gliding speed of migrating birds: slow and safe or fast and risky? *Ecol Lett*, 17, 670-679.

[31] Cochrane, J. H. (1999) MacCready theory with uncertain lift and limited altitude, *Technical Soaring*, 23:88-96.

[32] Grotzbach, G. (1983) Spatial resolution requirements for direct numerical simulation of the Rayleigh-Bénard convection *J. Comp. Phys*, 49, 241-264.

[33] Shamoun-Baranes, J., Leshem, Y., Yom-tov, Y. & Liechti, O., Differential Use of Thermal Convection by Soaring Birds Over Central Israel, The Condor, 105(2): 208-218, 2003.

[34] Weimerskirsch, H., Bishop, C., Jeanniard-du-Dot, T., Prudor, A. & Sachs, G., Frigate birds track atmospheric conditions over months-long transoceanic flights, Science, 353(6294): 74-78, 2016.

[35] Silver, D., Schrittwieser J., Simonyan K., Antonoglou I., Huang A., Guez A., Hubert T., Baker L., Lai M., Bolton A., Chen Y., Lillicrap T., Hui F., Sifre L., van den Driessche G., Graepel T., & Hassabis D., Mastering the game of Go without human knowledge, *Nature*, 550:354-359, 2017.

[36] Mnih, V., Kavukcuoglu K., Silver D., Rusu A. A., Veness J., Bellemare M. G., Graves A., Riedmiller M., Fidjeland A. K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S. & Hassabis D., Human-level control through deep reinforcement learning, *Nature*, 518:529533, 2014.

[37] Kim, H. J., Jordan, M. I., Sastry, S., & Ng, A., Autonomous Helicopter Flight via Reinforcement Learning, Advances in Neural Information Processing Systems, 16, 2003.

[38] Levine, S., Finn, C., Darrell, T. & Abbeel, P., End-to-End Training of Deep Visuomotor Policies, Journal of Machine Learning Research, 17:1-40, 2016.

[39] Edwards, D. J., Implementation Details and Flight Test Results of an Autonomous Soaring Controller, AIAA Guidance, Navigation and Control Conference and Exhibit, 2008.

[40] Edwards, D. J., Autonomous Soaring: The Montague Cross Country Challenge Doctorate theses, North Carolina State University, Aerospace Engineering, Raleigh, North Carolina, 2010.

[41] Chung J. J., Lawrance, N. R. J. & Sukkarieh, S, Learning to soar: Resource-constrained exploration in reinforcement learning, The International Journal of Robotics Research, 34(2):158-172, 2014.

[42] Reddy, G., Celani, A., Sejnowski, T. & Vergassola, M., Learning to soar in turbulent environments, Proc. Natl. Acad. Sci., 113(33): E4877-E4884, 2016.

[43] Yeung, P. K. & Pope, S. B., Lagrangian statistics from direct numerical simulations of isotropic turbulence, J. Fluid. Mech., 207:531-586, 1989.

[44] Voth, G. A., La Porta, A., Crawford, A. M., Alexander, J., & Bodenschatz, E., Measurement of particle accelerations in fully developed turbulence, J. Fluid. Mech., 469:121-160, 2002.

[45] Tennekes, H. & Lumley, J. L., A first course in turbulence, MIT Press, 1972.

[46] Ng, A. Y., Harada, D., & Russell, S. J., Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping, Proc. of the 16th International Conference on Machine

[47] ArduPilot, www.ardupilot.org, (2018).

[48] R. von Mises, *Theory of Flight*, Dover Publications, 1st edition, 1959.

[49] R. Stengel, *Flight Dynamics*, Princeton University Press, 1st edition, 2004.

[50] H. C. Berg & E. M. Purcell, Physics of chemoreception, *Biophysical Journal*, 20(2): 193219, 1977.

[51] I. S. Gradshteyn & I. M. Ryzhik, *Table of Integrals, Series, and Products*, ed. D. Zwillinger, Academic Press, 8th edition, 2014.