Glider soaring via reinforcement learning in the field

Gautam Reddy, Jérôme Wong Ng, Antonio Celani, Terrence J Sejnowski, Massimo Vergassola

Soaring through reinforcement learning in the field

Gautam Reddy[1]*, Jerome Wong Ng[1]*, Antonio Celani[2], Terrence J. Sejnowski[3,4], Massimo Vergassola[1]

[1]Department of Physics, University of California San Diego, La Jolla, USA, [2]The Abdus Salam International Center for Theoretical Physics, I-34014 Trieste, Italy, [3]The Salk Institute for Biological Studies, La Jolla, USA and [4]Division of Biological Sciences, University of California, San Diego, La Jolla, USA.

*These authors contributed equally to this work.

Abstract

Soaring birds often rely on ascending thermal plumes in the atmosphere as they search for prey or migrate across large distances[1-4]. 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[5-6]. How soaring birds find and navigate thermals within this complex landscape is unknown. Reinforcement learning[7] 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.

Main

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[8] and Go[9], visual-based video game play[10] and robotics[11,12]. 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. 11 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[13-16], which is unrealistic, or have applied learning methods in highly simplified simulated settings[17-19].

Using the reinforcement learning framework[7], 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[20]. Two cues emerged from our screening: (1) the vertical wind acceleration ($\boldsymbol{a}_z$) along the glider’s 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 in ref. 20 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.

To learn to soar in the field, we used a glider (of two-meter wingspan) with autonomous soaring capabilities (Figures 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 $\boldsymbol{a}_z$ by the derivative of the vertical ground velocity ($\boldsymbol{u}_z$), 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 ($\boldsymbol{w}_z$), and $\boldsymbol{a}_z$ 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 1C-D, together with traces of the vertical wind velocity, $\boldsymbol{w}_z$, μ and unbiased estimates of $\boldsymbol{a}_z$.

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 the Supplementary Information, 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 (Supplementary Information). 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[7]. 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. 21), which is of the order of minutes in the atmosphere. Conversely, the correlation timescale of accelerations is controlled by the smallest timescale[21-23] (the dissipation timescale in Fig. 7 of ref. 21). This is estimated to be only a fraction of a second, which is much smaller than the time interval between successive actions. Note that 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[23]. 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 2A), calculated using value iteration methods (Methods).

The navigational strategy derived at the end of the training period is presented in Figure 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[24] -- 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 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^{o} and -15^{o}. Importantly, the learned leftward/rightward bias in bank angle on encountering a positive/negative torque validates our estimation procedure for ω.

In Figure 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 Extended Data Figure 1). On a day with strong atmospheric convection, the time spent aloft is limited only by visibility and the receiver's 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 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 (Methods). 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 (Methods). Results shown in Figure 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.

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[25], 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[26-28]? This calls for the development of effective methods for identifying areas of strong updraft based on mechanical and visual cues. Such methods, coupled with our current work, pave the way towards a better understanding of how birds migrate and the development of autonomous vehicles that can extensively fly with minimal energy cost.

Main Figure Legends

Figure 1: Soaring in the field using turbulent navigational cues. (a) A trajectory of our glider soaring in Poway, California. (b) A cartoon of the glider showing the available navigational cues -- gradients in vertical wind velocities along the trajectory and across its wings, which generate a vertical wind acceleration $\boldsymbol{a}_z$ and a roll-wise torque ω respectively. (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 panel (c). The ω (solid, green) is estimated from the small deviations of the measured bank angle (solid, blue) from the expected bank angle (dashed, orange) after accounting for other effects (Methods).

Figure 2: Convergence of the learning algorithm and the learned thermalling strategy. (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, obtained by bootstrapping from the experiences accumulated up to that point. (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, as shown in the legend) to be taken when the glider observes a particular ($\boldsymbol{a}_z$,ω) pair and is banked at μ. Combined symbols depict pairs of actions that are equally rewarding. Note that a positive ω corresponds to a higher vertical wind velocity on the left (right) wing of the glider and a positive (negative) μ corresponds to turning right (left) w.r.t the glider's heading.

Figure 3: Performance of the learned strategy and its dependence on the wingspan. (a) A 12-minute-long trajectory of the glider executing the learned thermalling strategy in the

field, colored by the vertical ground velocity $\boldsymbol{u}_z$ at each instant. (b) Experimentally measured climb rate of a control random policy (black dots) is compared against the learned strategy (red dots) over repeated 3-minute trials in the field. Each dot represents the average climb rate in a single trial. A few outliers are not shown to restrict the range of the axis. (c) Estimated signal-to-noise ratio (SNR) in ω and $\boldsymbol{a}_z$ estimation *vs* wingspan (l) shown in green and red respectively. The SNR for ω estimation is plotted in log-log scale (inset) to highlight the weak $l^{1/6}$ scaling. (d) The mean climb rate for the learned strategy is compared for different wingspans (red dots) in simulations of a glider soaring in the convective boundary layer. For comparison we show the mean climb rates for a random policy and a strategy that uses $\boldsymbol{a}_z$ only (Methods). Error bars represent s.e.m.

Methods

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[29]. 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 Extended Data Figure 2 for the definitions of the angles). Throughout the paper, we use $\mu > 0$ when the plane is banked to the right and $\varphi > 0$ 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 Extended Data Table 1 for parameter values) such that:

$$\tau \frac{d\varphi}{dt} = \phi_d - \phi \qquad (1)$$

$$\tau \frac{d\mu}{dt} = \mu_d - \mu \qquad (2)$$

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:

$$\mu_d(t) = \mu_i + (\mu_f - \mu_i)(t + \tau)/t_a \qquad (3)$$

Estimation of the vertical wind acceleration. The vertical wind acceleration $\boldsymbol{a}_z$ is defined as:

$$\boldsymbol{a}_z \equiv \frac{dw_z}{dt} = \frac{d}{dt}(u_z - v_z) \qquad (4)$$

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 (Extended Data Figure 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 the Supplementary Information, 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:

$$\Delta v_z = -V(\Delta\alpha - \Delta\phi) \qquad (5)$$

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

$$\Delta\alpha \approx (\alpha_0 - \alpha_i)(\frac{1}{\cos\mu} - 1) \quad (6)$$

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 (α_0 - α_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 (Extended Data Figure 4).

Estimation of 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). We can then write an effective equation,

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau} + \omega(t) + \omega_{aero}(t) \qquad (7)$$

where $\omega(t)$ and $\omega_{aero}(t)$ are contributions to the roll-wise angular velocity due to the wind and aerodynamic effects respectively. We empirically find four major contributions to ω_{aero}: (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 when rolling at low airspeeds. We quantify the contributions from the four effects and model their dependence on the bank angle (see Supplementary Information for more details on modeling and calibration). A estimate of ω is then obtained as:

$$\omega = \frac{d\mu}{dt} - \frac{\mu_d - \mu}{\tau} - \omega_{aero} \qquad (8)$$

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

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 ref. 20. 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°, $\pm 15^\circ$, $\pm 30^\circ$, 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 × 3 × 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 ref. 20. In the Supplementary Information, 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.

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[7] 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 ref. 20.

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 Supplementary Information) 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 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 μ → −μ. This transforms a state as ($\boldsymbol{a}_z$,ω,μ) → ($\boldsymbol{a}_z$,-ω,-μ) 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

$$T^{sym} = \frac{T^{+} + T^{-}}{2} \qquad (9)$$

$$R^{sym} = \frac{R^{+} + R^{-}}{2} \qquad (10)$$

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 2b. To conveniently obtain the policy that uses only $\boldsymbol{a}_z$ (Figure 3d), the above procedure is repeated with the threshold for ω (K_ω) set to infinity.

Testing the performance of the learned policy in the field. To obtain the data shown in Figure 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 Extended Data Figure 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.

Testing the performance for different wingspans in simulations. Soaring performance is analyzed in simulations similar to those developed in ref. 20 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[30] and the mean velocity profile in the convective boundary layer[5], as described in the SI of ref. 20. Stationary updrafts and downdrafts of Gaussian shape are placed on a staggered lattice of spacing ~125m on top of the fluctuating velocity field. Specifically, their contribution to the vertical wind velocity at position *r* is given by

$$w_z = \pm W e^{-\frac{(r_\perp - r_\perp^0)^2}{2R^2}} \quad (11)$$

where $r^0_\perp$ 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°, $\pm 15^\circ$, $\pm 30^\circ$ in a time interval t_a, corresponding to the time step between actions. The glider's 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.

Estimation of the noise in gradient sensing due to atmospheric turbulence. The 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 the current $\boldsymbol{a}_z$ and ω. We estimate the magnitude of turbulent 'noise' that unavoidably accompanies gradient sensing. Intuitively, turbulent fluctuations in the atmospheric boundary layer (ABL) are made up of eddies of different length scales, with the largest being the size of the height of the ABL. Energy is transferred from larger, stronger eddies to smaller, weaker eddies, and eventually dissipates at the centimeter scale due to viscosity in the bulk and the boundaries. In the Supplementary Information, we present an explicit calculation of the signal to noise ratio for ω estimation taking into account the effect of turbulent eddies on the statistics of noise. Below, we give simple scaling arguments and refer to the Supplementary Information for further details.

A glider moving at an airspeed V and integrating over a time scale T averages $\boldsymbol{a}_z$ over a length VT. For V much larger than the velocity scale of the eddies, which is typically the case, the decorrelation of wind velocities is due to the glider's motion; the eddies themselves can be considered to be frozen in time. The magnitude of the spatial fluctuations across the eddy of this size scales according to the Richardson-Kolmogorov law[30] as ~ $(VT)^{1/3}$. The mean gradient signal when going up the gradient is ~ (VT); the resultant signal to noise ratio in $\boldsymbol{a}_z$ scales as $(VT)^{2/3}$.

Similar arguments are applicable for ω measurements. In this case, the signal to noise ratio has an additional dependence on the wingspan *l*. The dominant contribution to the noise comes from eddies of size *l*, whose strength scales as $l^{1/3}$. As the glider moves a distance *VT*, for $l \ll VT$, it traverses *VT/l* distinct eddies of size *l*. Consequently, the noise is averaged out by a factor $(VT/l)^{-1/2}$, corresponding to the *VT/l* independent measurements. Multiplying these two factors, the averaged noise is ~ $l^{5/6}(VT)^{-1/2}$. Since the mean gradient (i.e., the signal) is ~ *l*, the signal to noise ratio is then ~ $l^{1/6}(VT)^{1/2}$.

From the above arguments and dimensional considerations, we get order-of-magnitude estimates of the SNR for $\boldsymbol{a}_z$ and ω estimation:

$$SNR\ (a_z)\ \sim \frac{WV^{2/3}T^{2/3}L^{1/3}}{wR} \qquad (12)$$

$$SNR\ (\omega)\ \sim \frac{WV^{1/2}T^{1/2}l^{1/6}L^{1/3}}{wR} \qquad (13)$$

where *W* is the strength of the thermal, *R* is its radius, *w* is the magnitude of turbulent vertical wind velocity fluctuations and *L* is the length scale of the ABL. For the SNR estimates presented in the text, we use *W* = 2m/s, *R* = 50m, *l* = 2m, *V* = 8m/s, *T* = 3 s, *L* = 1 km. The values of *V* and *T* correspond to the airspeed of the glider in experiments and the time scale between actions during learning respectively.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Newton I., *Migration Ecology of Soaring Birds*, Elsevier, 1st edition, 2008.
2. Shamoun-Baranes, J., Leshem, Y., Yom-tov, Y. and Liechti, O., Differential Use of Thermal Convection by Soaring Birds Over Central Israel, *The Condor,* 105(2): 208-218, 2003.
3. Weimerskirsch, H., Bishop, C., Jeanniard-du-Dot, T., Prudor, A. and Sachs, G., Frigate birds track atmospheric conditions over months-long transoceanic flights, *Science*, 353(6294): 74-78, 2016.
4. Pennycuick, C. J., Thermal Soaring Compared in Three Dissimilar Tropical Bird Species, *Fregata Magnificens, Pelecanus Occidentals* and *Coragyps Atratus*, *J. Exp. Biol.,* 102:307-325, 1983.
5. Garrat, J. R., *The Atmospheric Boundary Layer*, Cambridge Atmospheric and Space Science Series, Cambridge University Press, 1994.
6. Lenschow, D. H., & Stephens, P. L., The role of thermals in the atmospheric boundary layer, *Boundary-Layer Meteorology*, 19:509-532, 1980.
7. Sutton, R. S., & Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, 1st edition, 1998.
8. Tesauro, G., Temporal difference learning and TD-Gammon, *Commun ACM*, 38: 58–68, 1995.
9. Silver, D. et al, Mastering the game of Go without human knowledge, *Nature*, 550:354-359, 2017.
10. Mnih, V. et al, Human-level control through deep reinforcement learning, *Nature*, 518:529–533, 2014.
11. Kim, H. J., Jordan, M. I., Sastry, S., and Ng, A., Autonomous Helicopter Flight via Reinforcement Learning, *Advances in Neural Information Processing Systems*, 16, 2003.

12. Levine, S., Finn, C., Darrell, T. and Abbeel, P., End-to-End Training of Deep Visuomotor Policies, *Journal of Machine Learning Research*, 17:1-40, 2016.
13. Allen, M. J. & Lin, V., Guidance and Control of an Autonomous Soaring UAV, *Proceedings of the AIAA Aerospace Sciences Meeting,* (American Institute of Aeronautics and Astronautics, Reston, VA), AIAA Paper 2007-867.
14. Edwards, D. J., Implementation Details and Flight Test Results of an Autonomous Soaring Controller, AIAA Guidance, Navigation and Control Conference and Exhibit, 2008.
15. Edwards, D. J., Autonomous Soaring: The Montague Cross Country Challenge Doctorate theses, North Carolina State University, Aerospace Engineering, Raleigh, North Carolina, 2010.
16. Ákos, Z., Nagy, M., Leven, S., and Vicsek, T., Thermal soaring flight of birds and unmanned aerial vehicles. *Bioinspiration & Biomimetics*, 5(4), 045003, 2010.
17. Doncieux, S., Mouret, J. B. and Meyer J-A., Soaring behaviors in UAVs : 'animat' design methodology and current results. In *7th European Micro Air Vehicle Conference (MAV07),* Toulouse, 2007.
18. Wharington, J. and Herszberg, I., Control of a high endurance unmanned aerial ve-hicle, *Proceedings of the 21st Congress of International Council of the Aeronautical Sciences* (International Council of the Aeronautical Sciences, Bonn, Germany), Paper 98-3.7.1, 1998.
19. Chung J. J., Lawrance, N. R. J. and Sukkarieh, S, Learning to soar: Resource-constrained exploration in reinforcement learning, *The International Journal of Robotics Research,* 34(2):158-172, 2014.
20. Reddy, G., Celani, A., Sejnowski, T. and Vergassola, M., Learning to soar in turbulent environments, *Proc. Natl. Acad. Sci.*, 113(33): E4877-E4884, 2016.
21. Yeung, P. K. and Pope, S. B., Lagrangian statistics from direct numerical simulations of isotropic turbulence, *J. Fluid. Mech.,* 207:531-586, 1989.

22. Voth, G. A., La Porta, A., Crawford, A. M., Alexander, J., and Bodenschatz, E., Measurement of particle accelerations in fully developed turbulence, *J. Fluid. Mech.,* 469:121-160, 2002.
23. Tennekes, H. and Lumley, J. L., *A first course in turbulence,* MIT Press, 1972.
24. Reichmann, H., *Cross-Country Soaring*, Thomson Publications, Santa Monica, CA, 1988.
25. Ng, A. Y., Harada, D., and Russell, S. J., Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping, *Proc. of the 16th International Conference on Machine Learning*, P278-287, 1999.
26. MacCready, P. B. J., Optimum airspeed selector, *Soaring*, 1958(1): 10–11, 1958.
27. Horvitz, N. et al., The gliding speed of migrating birds: Slow and safe or fast and risky? *Ecology Letters*, 17(6):670–679, 2014.
28. Cochrane, J. H., MacCready theory with uncertain lift and limited altitude, *Technical Soaring*, 23:88–96, 1999.
29. ArduPilot, www.ardupilot.org, (2018).
30. Frisch, U., *Turbulence: The Legacy of A. N. Kolmogorov*, Cambridge University Press, 1995.

Supplementary Information

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

This work was supported by Simons Foundation Grant 340106 (to M.V.).

Author Contributions

All authors were involved in designing the study and drafting the final manuscript. G.R. and J.W.N. performed the experiments and analyzed the data. G.R., A.C. and M.V. contributed to the theoretical results.

Author Information

Reprints and permissions information is available at www.nature.com/reprints.

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to M.V (massimo@physics.ucsd.edu).

Extended Data Legends

Extended Data Table 1: Parameter values.

Extended Data Figure 1: Sample trajectories obtained in the field. The three-dimensional view and top view of the glider's trajectory as it executes the learned thermalling strategy (labeled 's') or a random policy that takes actions with equal probability (labeled 'r'). The trajectories are colored with the instantaneous vertical ground velocity ($\boldsymbol{u}_z$). The green (red) dot shows the start (end) point of the trajectory. Trajectories s1, s2 and r1 last for 3 minutes each, whereas s3 lasts for ~8 minutes.

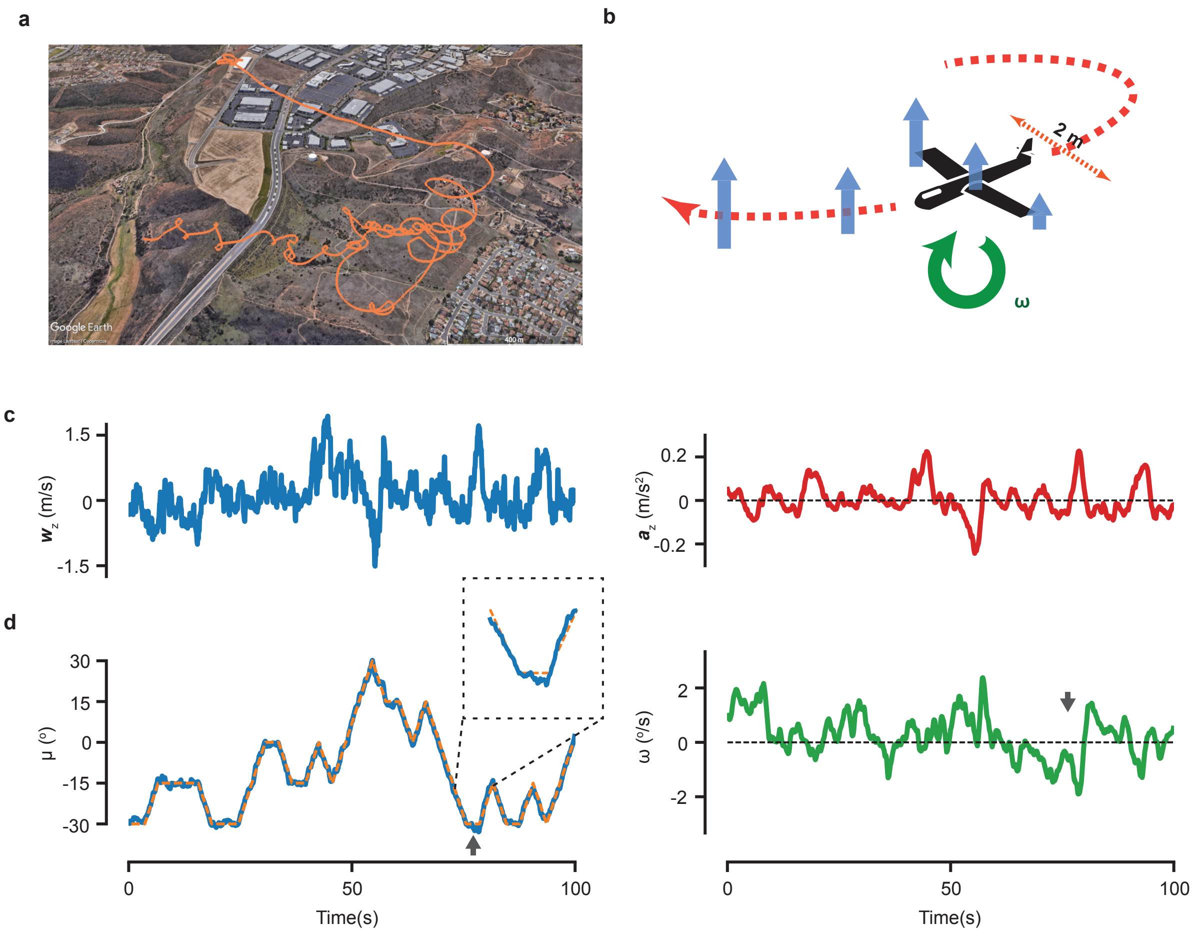
Extended Data Figure 2: Force-body diagram of a glider. The forces on a glider and the definitions of the various angles that determine the glider's motion.

Extended Data Figure 3: Modeling the longitudinal motion of the glider. (a) A sample trajectory of a glider's pitch and its vertical velocity w.r.t ground $\boldsymbol{u}_z$ in a case where the feedback control over the pitch is reduced in order to exaggerate the pitch oscillations. The blue line shows the measured $\boldsymbol{u}_z$ and the orange line is $\boldsymbol{u}_z$ obtained after subtracting the contributions from longitudinal motions of the glider (see Supplementary Information). (b) The blue line shows the average change in $\boldsymbol{u}_z$ when a particular action is taken (labeled above each panel), averaged over *n* three-second intervals. The 13 panels correspond to the 13 possible bank angle changes from the angles 0^o, $\pm 15^o$, and $\pm 30^o$ by increasing, decreasing the bank angle by 15^o or keeping the same angle. The green, dashed line shows the prediction from the model whereas the orange line is the estimated $\boldsymbol{w}_z$. The axis on the right shows the averaged pitch as a red, dashed line.

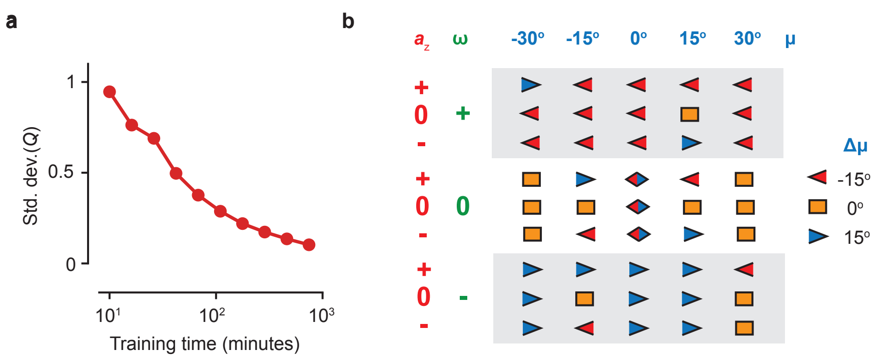
Extended Data Figure 4: The estimated vertical wind acceleration is unbiased after accounting for the glider's longitudinal motion. (a) The averaged vertical wind acceleration, $\boldsymbol{a}_z$ in units of its standard deviation $\boldsymbol{a}_z$, plotted as in Extended Data Figure 3b, is shown in orange with (blue line) and without (orange line) accounting for the glider's longitudinal motions. The axis on the right shows the airspeed as a green, dashed line. (b) The PDFs (probability density functions) of $\boldsymbol{a}_z$ for the different bank angle changes. The black, dashed line shows the median.

Extended Data Figure 5: The estimated roll-wise torque is unbiased after accounting for the effects of feedback control and glider aerodynamics. (a) The averaged evolution of the bank angle shown as in Extended Data Figure 3b. The blue line shows the measured bank angle and the dashed, orange line shows the best-fit line obtained from simultaneously fitting the 13 blue curves to the prediction (see Supplementary Information). (b) The PDFs (probability density functions) of the roll-wise torque ω (in units of its standard deviation) for the different bank angle changes. The black, dashed line shows the median value.

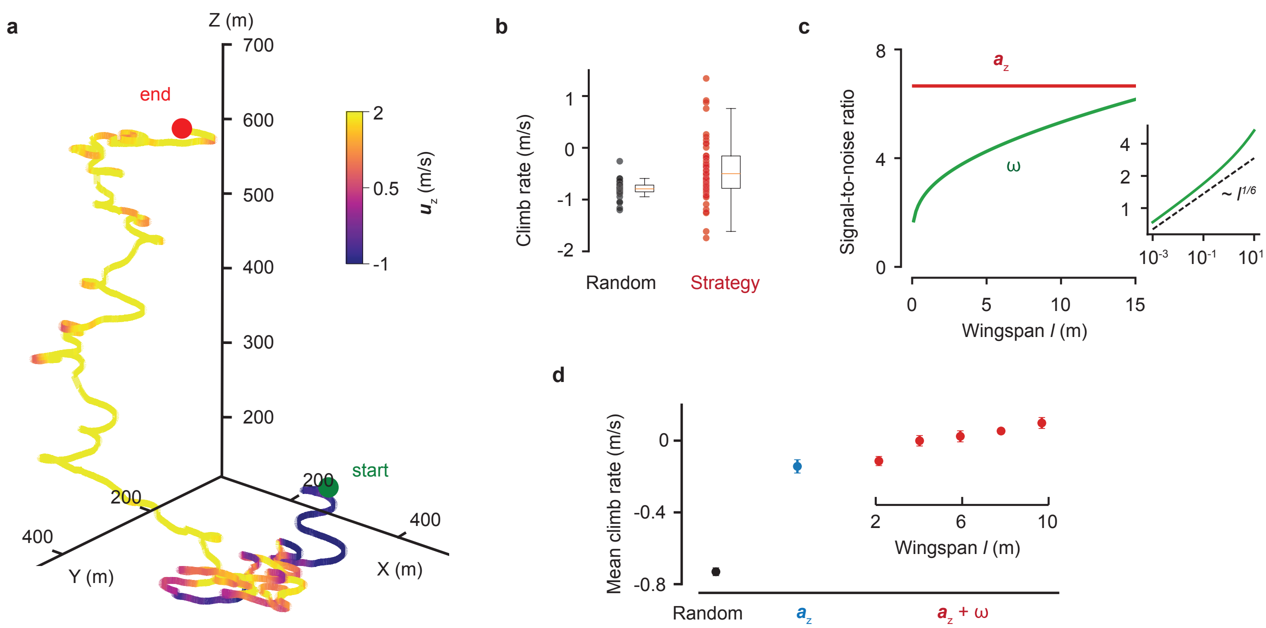
Extended Data Figure 6: The distribution of the strength of vertical currents observed in the field. The root-mean-square vertical wind velocity measured in the field is pooled from ~240 3-minute trials collected over 9 days. The dashed, red line shows the threshold criterion imposed when measuring the performance of the strategy in the field (see Methods).



a
b
Google Earth
2 m
ω
c
1.5
0
-1.5
w_z (m/s)
0.2
0
-0.2
a_z (m/s^2)
d
30
15
0
-15
-30
μ (°)
2
0
-2
ω (°/s)
0
50
100
Time(s)
0
50
100
Time(s)



a
b
1
0.5
0
Std. dev.(Q)
10^1
10^2
10^3
Training time (minutes)
a_z
ω
-30°
-15°
0°
15°
30°
μ
+
0
-
Δμ
-15°
0°
15°


a
Z (m)
700
600
500
400
300
200
end
start
200
400
200
400
Y (m)
X (m)
2
0.5
-1
u_z (m/s)
b
Climb rate (m/s)
1
0
-1
-2
Random
Strategy
c
Signal-to-noise ratio
8
4
0
a_z
ω
0
5
10
15
Wingspan l (m)
4
2
1
10^{-3}
10^{-1}
10^{1}
$\sim l^{1/6}$
d
Mean climb rate (m/s)
0
-0.4
-0.8
2
6
10
Wingspan l (m)
Random
a_z
a_z + ω