北京航空航天大學
BEIHANG UNIVERSITY

LITERATURE READING

Glider soaring via reinforcement learning in the field

Nature 2018.

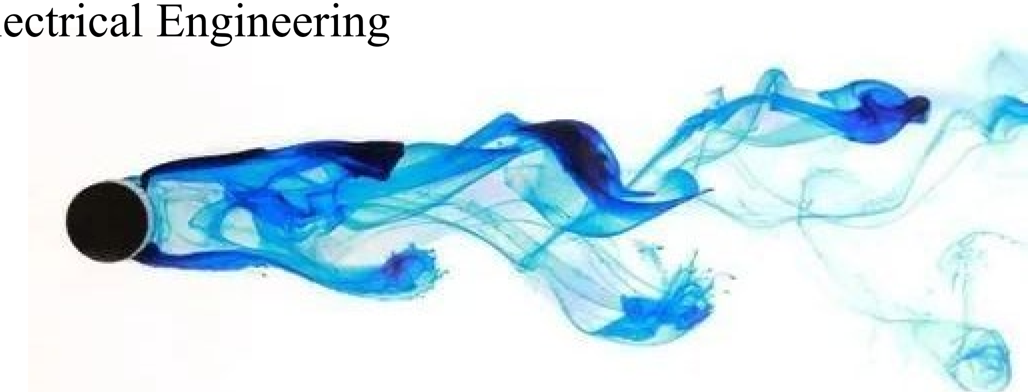
Learning to soar in turbulent environments

Proceedings of the National Academy of Sciences 2016.

Jinjie LI

School of Automation Science and Electrical Engineering
Beihang University

March. 26, 2021



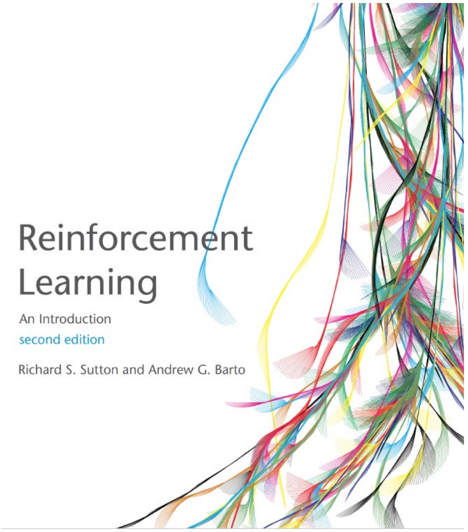
Outline

- ❑ Background
- ❑ Challenge
- ❑ Solutions
- ❑ Discussion
- ❑ My idea
- ❑ Top journals and conferences in the field of robotics

Outline

- ❑ Background
- ❑ Challenge
- ❑ Solutions
- ❑ Discussion
- ❑ My idea
- ❑ Top journals and conferences in the field of robotics

Background



16 Applications and Case Studies 421
16.1 TD-Gammon 421
16.2 Samuel's Checkers Player 426
16.3 Watson's Daily-Double Wagering 429
16.4 Optimizing Memory Control 432
16.5 Human-level Video Game Play 436
16.6 Mastering the Game of Go 441
16.6.1 AlphaGo 444
16.6.2 AlphaGo Zero 447
16.7 Personalized Web Services 450
16.8 Thermal Soaring 453

PNAS

Learning to soar in turbulent environments

Gautam Reddy[a], Antonio Celani[b], Terrence J. Sejnowski[c,d,1], and Massimo Vergassola[a]

[a]Department of Physics, University of California, San Diego, La Jolla, CA 92093; [b]The Abdus Salam International Center for Theoretical Physics, I-34014 Trieste, Italy; [c]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037; and [d]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093

Contributed by Terrence J. Sejnowski, April 28, 2016 (ser

Birds and gliders exploit warm, rising atmospheric cu to reach heights comparable to low-lying clouds expenditure of energy. This strategy of flight (th

LETTER

https://doi.org/10.1038/s41586-018-0533-0

Glider soaring via reinforcement learning in the field

Gautam Reddy[1,5], Jerome Wong-Ng[1,5], Antonio Celani[2], Terrence J. Sejnowski[3,4] & Massimo Vergassola[1]*

Soaring birds often rely on ascending thermal plumes (thermals) is unrealistic, or have applied learning methods in highly simplified

- ❑ SARSA. 这合理吗?

Background

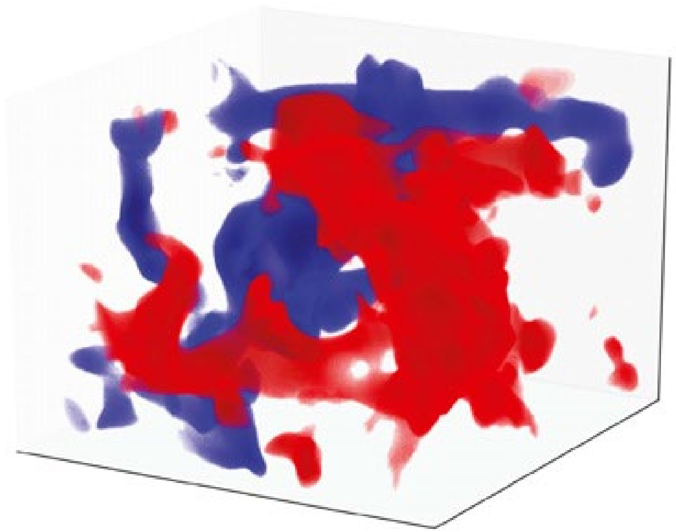


Fig. 1 风速流场，红色上升，蓝色下降

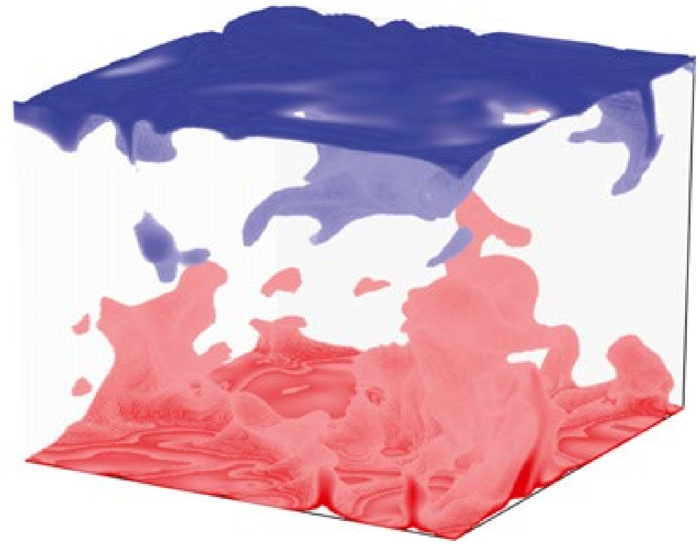


Fig. 2 温度分布，红色高温，蓝色低温

- 滑翔 (soaring)。目标：保持在热气流的中心
- **Challenge**：不可复现，难以建模，湍流干扰极大
- 问题：鸟类是如何感知这些气流的？什么是线索？

--- “Learning to soar in turbulent environments”

Solution

核心想法：使用强化学习，通过结果的一致性推断过程的一致性

基本原则：最小化控制所需的生物或电子传感器---实物实验的考量

核心方法：提炼出在以往文献中提出的可能会产生影响的因素：

垂直方向上的风速u_z，垂直方向上的风的加速度a_z，滚转力矩τ，局部的温度θ，以及它们的16种组合。目前的Bank Angle也是状态。

在仿真环境中，使用RL算法去不断的训练，找到结果较好的case对应的状态(state)。

算法：On-line, on-policy: SARSA，Tabular (Tile Coding)

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
    Initialize S
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Loop for each step of episode:
        Take action A, observe R, S'
        Choose A' from S' using policy derived from Q (e.g., ε-greedy)
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
        $S \leftarrow S'$; $A \leftarrow A'$;
    until S is terminal

WHY NOT DRL?

- 需要快速收敛
- 可解释性：需要通过观察Q表的变化，判断趋势

Solution

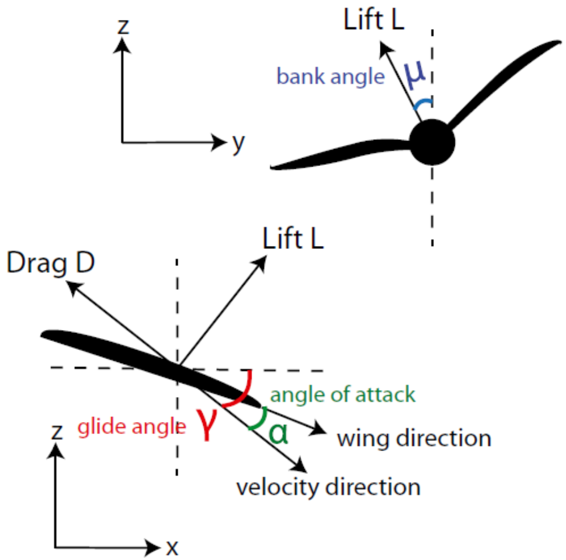
Markov Process：State，Action，Reward（实验过程介绍）

动作空间的选取：

bank angle（增加5°，不动，减少5°）

attack angle（增加2.5°，不动，减少2.5°）

共3^2 =9种组合。



奖励函数的选取：

1. 全局奖励，陷入稀疏奖励问题，用Eligibility traces也无法解决。

2. 最终选取：飞机落地给一个巨大的惩罚；每一步后$R = u_z + Ca_z$

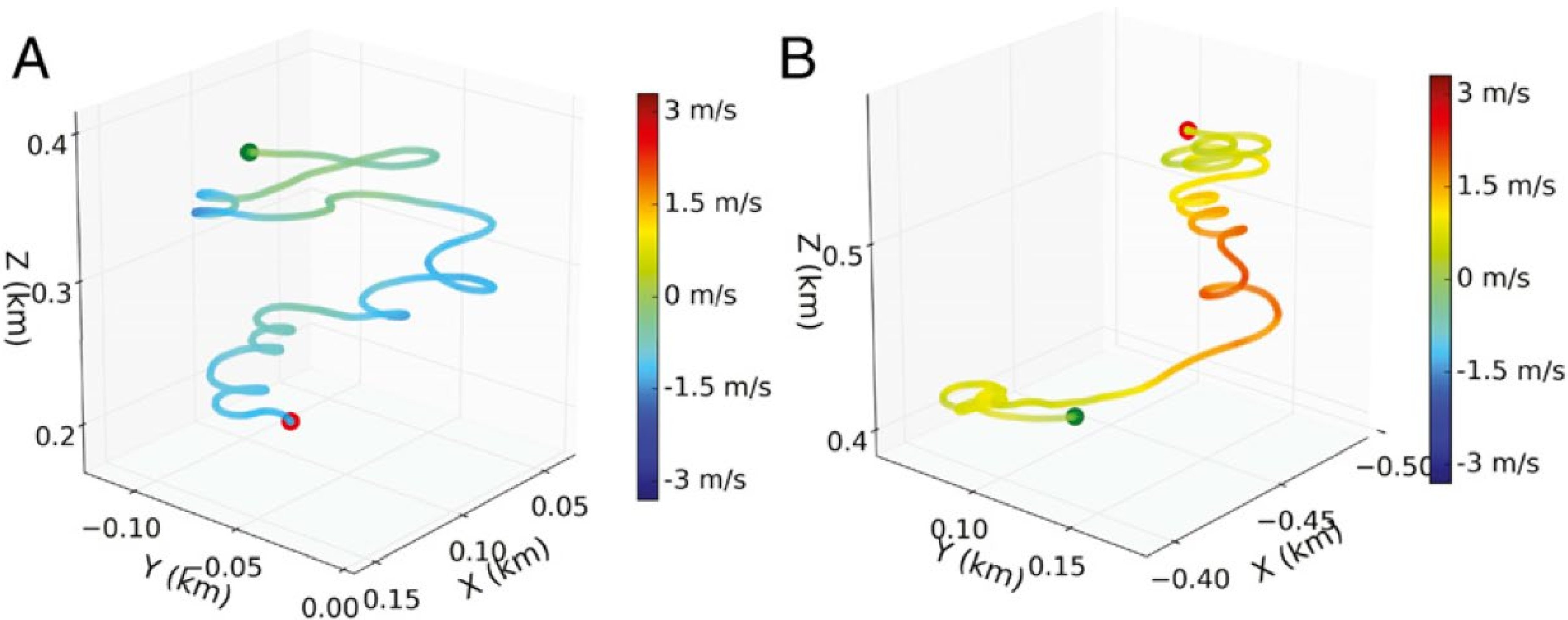
u_z是垂直向上的风速，a_z是垂直向上的风的加速度。我认为是观察Q表得出的。

其他信息：

- 湍流模型来自已知文献，分为强湍流和弱湍流两种环境去训练。
- 动作频率与训练周期：每次训练2.5min（birds 10min），动作执行周期1s（glider）

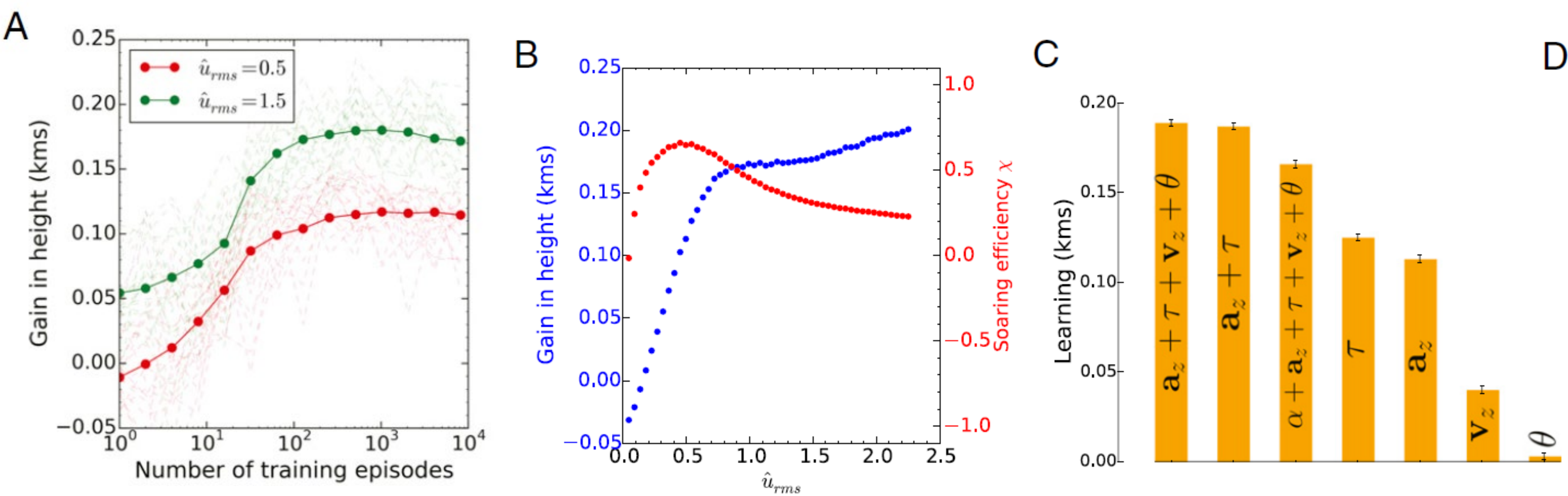
Solution

训练结果



Solution

训练结果

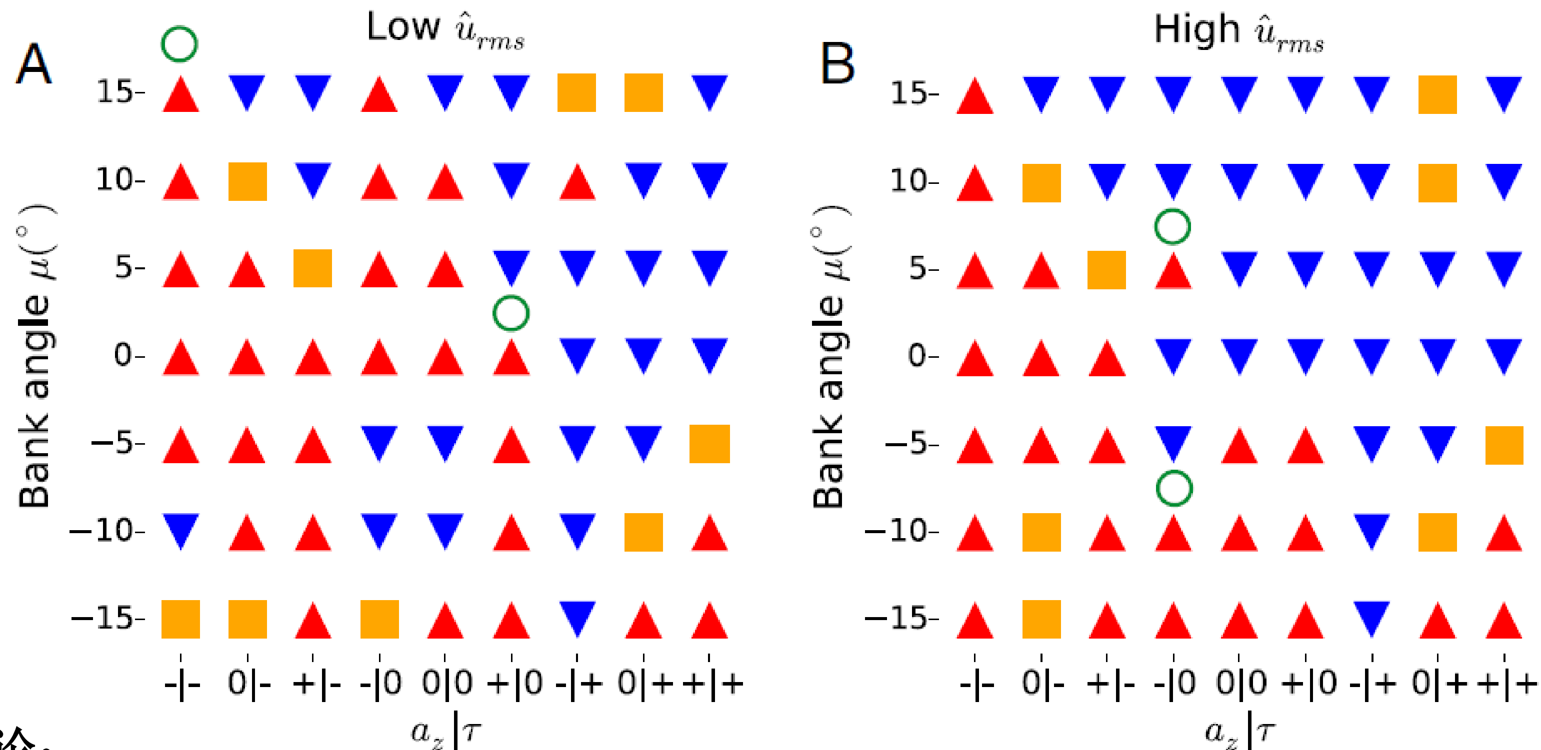


结论：

- 250次就收敛了。湍流越强，飞行高度越高，但找气流效率降低了。
- 垂直方向上的风的加速度a_z +滚转力矩τ，是**最能反应问题**的状态组合。

Solution

训练结果



结论：

- Tau为-，代表右翼向上速度大于左翼，向左滚转，要往右飞，bank angle增大。
- 在某个状态下的最优bank angle。
- 弱湍流环境，策略激进。强湍流环境，策略保守。非常精彩。

Challenge

- 如何在实物飞机上验证结论?

LETTER

https://doi.org/10.1038/s41586-018-0533-0

Glider soaring via reinforcement learning in the field

Gautam Reddy[1,5], Jerome Wong-Ng[1,5], Antonio Celani[2], Terrence J. Sejnowski[3,4] & Massimo Vergassola[1]*

核心想法：使用上一篇得到的结论设计RL过程，验证结论

- 问题：样本量极低，干扰极大，损害率极高

- 每天每时每刻的风强都是不一样的
- 如何估计状态？
- 需要精巧的实验设计收集样本，避免损害

Solution

Markov Process：State，Action，Reward （实验过程介绍）

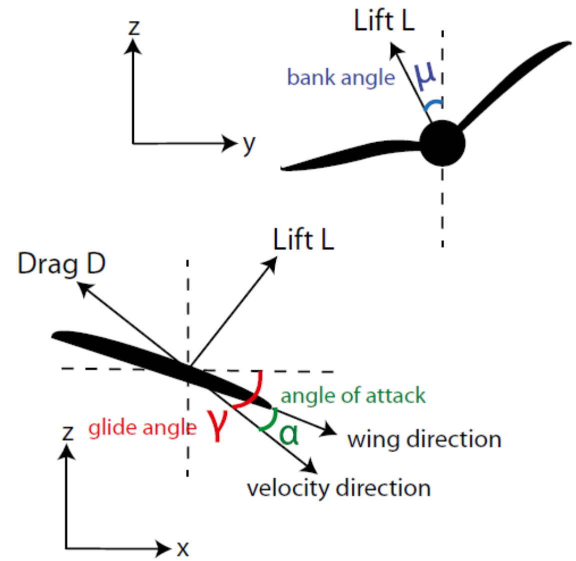
状态空间的选取：

垂直方向上的风的加速度a_z (+，0，-)

滚转力矩ω (+，0，-)

Bank angle (30, 15, 0, -15, -30)°

共3 × 3 × 5 =45个状态



动作空间的选取：

bank angle（增加15°，不动，减少15°）

奖励函数的选取：

垂直方向上的风的加速度a_z。这里可能考虑了估计准确性的问题。

算法的选择：

Off-line, offline-policy: Value Iteration，Tabular (Tile Coding)

WHY OFF-LINE?

- 在线收集数据，离线训练。

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

Solution

Markov Process：State，Action，Reward （实验过程介绍）

实验基本设定：

机型：Parkzone Radian Pro fix-wing， 2-m wing

自驾仪：Pixfalcon硬件，Ardupilot固件

传感器：GPS，磁罗盘，气压计，空速计，惯导

估计方法：EKF；控制方法：Proportional-Integral-Derivative

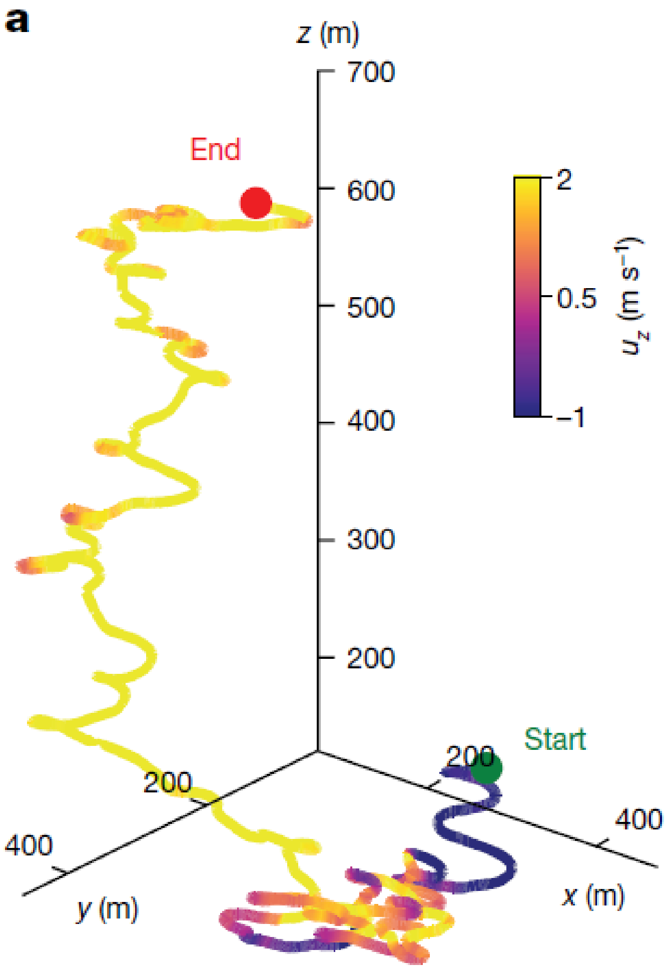
训练：

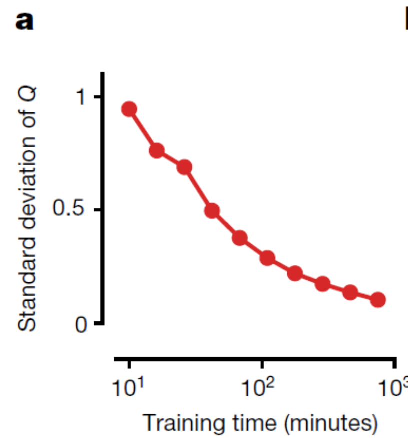
飞到250m的高度，开始训练。在前12天，采用完全随机策略，3s动作一次，一次飞行3min，训练12天。之后采用softmax policy训练3天。训练收集的数据会放进经验池中，离线训练。**使用Q-table的变化观察训练效果**。

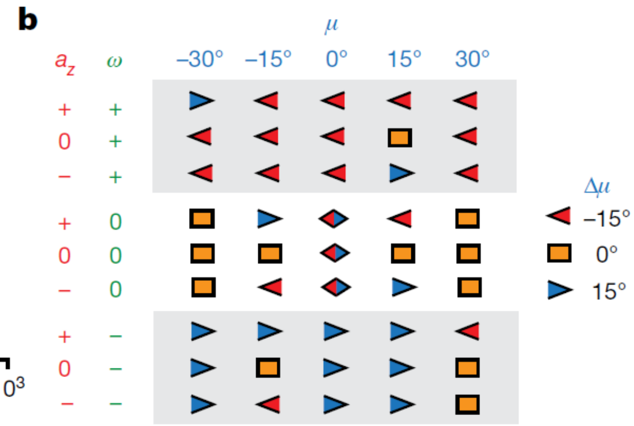
测试：

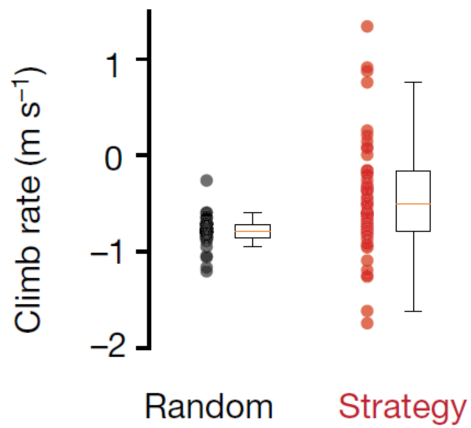
飞到250m高度，采用Q-table的策略，1.5s动作一次，飞行3分钟，看飞行高度。

Solution









结论：

- 成功学习到策略

Outline

- ❑ Background
- ❑ Challenge
- ❑ Solutions
- ❑ Discussion
- ❑ My idea
- ❑ Top journals and conferences in the field of robotics

Discussion

❑ 总结

- 状态未知、奖励不定，动作不定的问题。生物相关。强化学习。
- 仿真中通过**on-line**，**on-policy**算法进行状态量的选择，奖励函数的设定，动作的选取。
- 以上一点获得的RL设定为蓝本，通过**off-line**，**off-policy**算法进行实物训练
- 在以上基础上可以继续进行on-line学习。

❑ 启示

- 科研，是**算法引导问题，还是问题引导算法**?
- 不要忽视Tabular算法收敛快的特性。暴力训练一点都不优雅。
- 这两篇文章的研究模式，或许适合大多数**仿生决策问题**往机器人上的落地。
- 仿生是reward的重要来源。
- 极好的实验条件，巧妙的实物实验设计。

My idea

DJI FPV基本参数：

- 最大平飞速度140.4km/h，中国区是97.2km/h
- 最大下降速度和上升速度不限制
- 最大水平飞行加速度 0-100km/h，2s

现有方法

1. 无人机抓无人机

J. Rothe, M. Strohmeier and S. Montenegro, "A concept for catching drones with a net carried by cooperative UAVs," *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, Würzburg, Germany, 2019, pp. 126-132, doi: 10.1109/SSRR.2019.8848973.

Drone Catcher：https://dronecatcher.nl/#

2. 枪类、火箭筒：射程400m

My idea

My idea

Outline

- ❑ Background
- ❑ Challenge
- ❑ Solutions
- ❑ Discussion
- ❑ My idea
- ❑ Top journals and conferences in the field of robotics

“弄斧必到班门”

Journals:

- Nature, Science, PNAS
- Science Robotics
- IJRR (International Journal of Robotics Research)
- TR-O (IEEE Transactions on Robotics)

Conferences:

- RSS (Robotics: Science and Systems)
- CoRL (Conference on Robot Learning) 小众
- ICRA & IROS, 参差不齐

Thanks for your attention!

Q&A