Understanding biological plume tracking behavior using deep reinforcement-learning

Satpreet H. Singh1, Floris van Breugel2, Rajesh P. N. Rao1, and Bingni W. Brunton1

1University of Washington, Seattle, WA 98115
2University of Nevada, Reno, NV 89557
satsingh@uw.edu

Abstract
The ability to track odor plumes in dynamic environments is critical for flying insects following attractive odors to localize food or mates. This remarkable tracking behavior requires multimodal integration of odor, vision, and wind sensing, is robust to variations in plume statistics and wind speeds, and can often be performed over large distances. Therefore, it is challenging to study in confined experimental settings. Here we describe ongoing work to explore the space of policies effective to accomplish plume tracking, leveraging the reproducibility and interpretability of artificial agents trained in biologically motivated simulations. Specifically, we trained neural-network (NN) agents with deep reinforcement learning to locate the source of a patchy simulated plume, while varying their capacity to store past sensory stimuli. We analyzed the behavior of trained agents by inspecting successful trajectories. We then interrogated the input-output maps learned by the NNs, uncovering interpretable differences in control strategies introduced by varying sensory memory. We believe that our simulation-based approach can generate novel testable hypotheses to guide the development of targeted neuroethological experiments, as well as provide a pathway towards a mechanistic understanding of the key multimodal computations required for plume tracking.

Introduction
Flying insects often navigate dynamic environments to localize plume sources, a remarkable behavior that requires that they integrate mechanosensory, olfactory, and visual cues (Baker et al., 2018; Park et al., 2016; Cardé and Willis, 2008). While plume tracking is important to study since it is critical for survival, it has been difficult to study in controlled experimental settings due to the large spatial scales over which it is performed (tens of meters to ≈100 meters; see Wall and Perry 1987). Here we describe a simulation-centric approach to studying plume tracking, using artificial agents trained using reinforcement-learning (RL).

Neural-networks (NNs) have been broadly applied to model and understand different aspects of neural function and behavior, including visual (Kriegeskorte, 2015) and motor (Sussillo et al., 2015) systems (see also Kietzmann et al., 2019; Cichy and Kaiser 2019 for recent surveys). This literature has relied largely on supervised techniques that require labeled training data. In contrast, there has been relatively little work done in the RL setting, where an agent learns from interacting with its environment while only receiving supervision in the form of sparse rewards (Sutton and Barto, 2018). Recent algorithmic advances have made efficient training of NNs possible in the RL setting (Arulkumaran et al., 2017), resulting in their adoption in the neuroscience community for investigations into myriad phenomena such as grid-cell formation (Banino et al., 2018; Kanitscheider and Fiete, 2017), cortical meta-learning (Botvinick et al., 2019), and collective swimming (Verma et al., 2018). Classic RL (without NNs) has also been applied to understand bioinspired control of glider flight (Reddy et al., 2016) and flow navigation by microswimmers (Colabrese et al., 2017).

Here we present an approach for exploring hypotheses for control strategies for an agent solving a multimodal plume tracking task. To successfully complete the task, the agent must integrate sensory inputs corresponding to (relative) wind direction and odor concentration, and fly upwind such that it approaches the odor source and does not lose track of the plume. Using RL algorithms, we trained NNs with varying ability to model the history of sensory observations they encounter during a tracking episode. We analyze the differences in emergent behavior and control policies learned by our agents and explore their connections to biology.

Training plume tracking agents
Plume simulation Following the model introduced by (Farrell et al., 2002), our simulated plume consists of two parts: a constant-velocity fixed-direction wind field, and an odor source that emits odor packets with a Poisson birth-time distribution. After an odor packet is emitted, it is advected by the wind at each simulation step and translated by a small normally-distributed random perturbation in the crosswind direction, effectively performing a biased random walk downwind. Simulations are run at 100 frames/second.

Partially Observable Markov Decision Process (POMDP) To train agents using RL, we first define a POMDP as follows: State (observable) space is defined by the egocentric sensory observation triplet...
Figure 1: Example trajectories from agents with 0 (top), 5 (middle) and 10 (bottom) timesteps of memory: Plume shown in red. Crosshair indicates plume source. Arrows indicate $L \rightarrow R$ wind. Trajectories display an increasing ability of agent to make turns as memory capacity increases, including excursions outside of expanse of plume.

\[ S_o = [s_c, s_x, s_y] : s_c \in \mathbb{R}^+, s_x, s_y \in \mathbb{R} \] i.e. odor concentration and (x, y) coordinates of relative wind velocity as observed by the agent. Observables depend on the unobserved (hidden) state comprised of the agent’s current location and orientation in the plume. To introduce memory into this currently memoryless formulation, we augment the $S_o$ triplet with time-lagged triplets i.e. $S_o^{(0,1...L)} = [s_c^{(0)}, s_x^{(0)}, s_y^{(0)}, \ldots, s_c^{(L)}, s_x^{(L)}, s_y^{(L)}]$ for $L$ timesteps of lag. Action space is defined by the tuple: $A = [a_g, a_m] : a_g \in [-\pi, +\pi], a_m \in \mathbb{R}^+$ corresponding to how much the agent should turn and move forward. Rewards primarily depend on the agent’s location (hidden state): +100 if the agent reaches within a small fixed radius $r_1$ of the source, $-\epsilon$ per timestep taken during a training episode to encourage faster convergence, and as a form of reward shaping, a reward of $(r_{t_o} - r_{t_f})$ at the end of each episode corresponding to the reduction in the agent’s radial distance from the plume source between the start ($t_0$) and end ($t_f$) of the episode. Transitions deterministically update agent state obeying the system’s underlying physics. Agents start each episode at a random location and orientation within the plume. Episodes end when the agent reaches within $r_1$ of the source of the plume, or conversely, strays too far away from the plume, or exceeds a fixed time limit.

Deep Reinforcement Learning (DRL) We implemented the aforementioned POMDP environment using the OpenAI Gym (Brockman et al., 2016) and stable-baselines (Hill et al., 2018) libraries. Next, we trained feedforward NNs with 2 hidden-layers and $tanh$ nonlinearities for 500K steps using the Deep Deterministic Policy Gradient (Lillicrap et al., 2015) algorithm. Training produced an NN that implemented a control policy $\pi : S_o \rightarrow A$.

Analysis of agent behavior

We investigated the effect of different amounts of sensory memory on the behavior of trained agents. Figures 1 and 2 show example trajectories and learned control policies respectively for agents with 0, 5 and 10 timesteps of history. Additionally, we quantified the performance difference between agents by calculating the fraction of successful episodes over 500 evaluation episodes. These were 0.57, 0.96 and 0.87 for the agents with 0, 5 and 10 steps of memory respectively. In general, adding memory enables more complicated turning behaviors (Fig. 1), including excursions from the boundary of the plume that show indications of “casting” like behavior seen in moths and flies (Baker et al., 2018; Cardé and Willis, 2008; Budick and Dickinson, 2006).

Discussion

Our preliminary work demonstrates the importance of sensory memory for learning complex control strategies and successful completion of a multimodal plume tracking task. These initial results motivate various directions for extension, including towards more challenging environments that could involve larger spatial scales and nonstationary wind fields; and towards the use of recurrent NN models, that could learn the appropriate amount of sensory memory required from experience. Motivated by recent literature we also plan to investigate neural representations (Merel et al., 2019; Haesemeyer et al., 2019), and generalization (Hasson et al., 2020) between artificial and biological networks.

Acknowledgements: Funded by MURI AFOSR award FA9550-19-1-0386 (BWB) and Washington Research Foundation (BWB); NSF award EEC-1028725 (RPNR), a TWCF grant and a CJ and Elizabeth Hwang Professorship (RPNR)
References


