International School for Advanced Studies
(SISSA)

Search, navigation and foraging: an optimal decision-making perspective

Thesis for the Ph. D. in Physics and Chemistry of Biological Systems

Supervisor
Antonio CELANI

Candidate
Matteo ADORISIO

IX Cycle — Academic Year 2017/2018

Abstract

Behavior in its general form can be defined as a mapping between sensory inputs and a pattern of motor actions that are used to achieve a goal. Reinforcement learning in the last years emerged as a general framework to analyze behavior in its general definition. In this thesis exploiting the techniques of reinforcement learning we study several phenomena that can be classified as search, navigation and foraging behaviors. Regarding the search aspect we analyze random walks forced to reach a target in a confined region of the space. In this case we can solve analytically the problem that allows to find a very efficient way to generate such walks. The navigation problem is inspired by olfactory navigation in homing pigeons. In this case we propose an algorithm to navigate a noisy environment relying only on local signals. The foraging instead is analyzed starting from the observation that fossil traces show the evolution of foraging strategies towards highly compact and self-avoiding trajectories. We show how this optimal behavior can emerge in the reinforcement learning framework.

Contents

Chapter 1

Introduction

General framework and a synopsis for the impatient reader

The role of a paradigm in science is to provide the scientific community with a sort of guiding principle to approach unexplored problems. Marr's levels of analysis [1, 2] proposed in the 1970s is one of these paradigms and it provided the idea that the understanding of a complex information-processing system necessarily passes through different levels of investigation:

> **Computational level** closely linked to an abstract framework regarding what is the computational problem to be solved and why
>
> **Algorithmic level** related to what are the sets of rules solving the abstract problem stated in the previous level
>
> **Implementational level** describing how the rules of the algorithmic level are effectively implemented in the system under examination.

At that time D. Marr and T. Poggio were considering the brain and in particular the vision as the goal of their paradigmatic analysis but over the years this three-levels "manifesto" became a useful conceptual tool to investigate and understand something that we can generally call a *decision-making system* or an *agent* as it will be useful in the following.

To understand its range of influence let us consider a biomimetic example. Imagine to have to design an agent capable of localizing a source of chemical in a turbulent atmosphere. This is the general goal and it can be placed at the abstract level of computation. Among all the organisms capable of localizing odor sources without any doubt a moth is a very indicated organism to mimic. The very faint pheromone signal emitted upwind by the female is transported and distorted by turbulence, mixed with other odors but nonetheless can be used by the exquisitely sensitive olfactory system of the male to locate very quickly the source [3, 4]. At this stage

the agent we want to design and the organism taken as an example share the same abstract capabilities and purpose.

At the algorithmic level we would like to ask how the agent represents the input to describe the current situation and map it into an output command.
It is clear that this level is intimately related to behavior that in its general form can be defined as a mapping between sensory inputs and a pattern of motor actions which then are used to achieve a goal. Moths are known to proceed upwind by alternating extended zigzagging behavior and upwind straight lines, thought to correlate with low and high rates of pheromone detection.
Algorithmically a given behavior can be obtained as a byproduct of the fact that the goal has been quantitatively specified in a precise way (e.g. the case of Infotaxis[1] [5]) or just hard-coding fixed action patterns related to plausible biological assumptions (see [6] and references therein). A crucial aspect at this stage regards how the agent represents the surrounding environment. A general criticism raised for Infotaxis, for example, is that it is unlikely that simple insects have the capabilities required by the algorithm. Nonetheless this algorithm represented the starting point to inspire a lot of research about olfactory robots. This aspect brings us to the the third level.

The level of implementation addresses the step connecting the idea given by the algorithm to the actual way in which the instructions are implemented. Even if a moth could not have the sufficient cognitive substrate to support Infotaxis a robot with modern technology easily can, and in the last years growing interest has been devoted to the design of bio-inspired robots capable of searching in condition in which odors are very dilute (for a review see [7]).

One thing that is not made explicit in Marr's diagram but is somehow present at the intersection between the computational and algorithmic level is the idea of *learning*. In the past few decades in particular Reinforcement Learning (RL) [8] emerged as the natural framework to study situations in which an entity called an agent repeatedly interacts with an environment to achieve a goal. At this stage we purposely keep abstract the definitions of agents, tasks and environment. Learning by reinforcement has its origin in classical and operant conditioning experiment by I.Pavlov, B.F.Skinner and E.Thorndike [9] that were focusing on learning by trial and error in animal behavior. Now it encloses disciplines as varied as psychology, computer science, neuroscience and ethology. In its generality RL is at the same time a problem (i.e. the agent has a goal) and a class of solution methods (i.e. how the agent makes decisions to achieve the goal). The evolution of this research field can be described thinking about two tracks that were running in parallel. One track that originated naturally at the intersection of animal behavior and computer science giving rise to the quest for artificial intelligence [10]. On the other track the theory of *optimal control* became popular in the 1950s thanks to the work of Richard

[1] In Infotaxis moth-like trajectories are not obtained hard coding the actions but asking the agent to locate the source mathematically formulated as a local maximization of the expected rate of information gain about the source position.

Bellman [11, 12] that formalized the problem of designing a controller to maximize or minimize some proxy for the system's behavior over time. Bellman formalized this concept into what now is called the *Bellman equation* that at the beginning did not involve the concept of learning but then became a central object in modern RL. Indeed, the track of *learning* and the one of *optimal control* remained separated until it was realized that the Bellman equation could be cast into a particular algorithmic but still very general form and solved using what are now called temporal difference learning algorithms[8, 13]. This shift made *learning* and *optimal control* to intersect giving rise to a set of ideas and methods characterizing the modern theory of Reinforcement Learning.

The mathematical framework enclosing all the ideas about RL will be presented in the next chapter. Here instead, to navigate the extended landscape of tools and concepts used in Reinforcement Learning it is useful to refer to figure 1.1 that will visually guide us. Figure 1.1 shows the typical scheme of perception-action cycle typical of a Reinforcement Learning problem. An *agent* chooses *actions* to maximize a given performance measure using its own *perception* of the surrounding environment and receiving from it a *feedback* on the performance.

The goal of a decision-making system is to maximize the long term return, essentially the sum of rewards. Experiments on animal behavior suggest the reward as a prize, for example in terms of food, to reinforce particular good actions but in recent years the concept of reward acquired wider definitions. For example to highlight its abstract structure studies focusing on the goal of intrinsic motivation (e.g. "curiosity") suggest that agents have to maximize the reduction of uncertainty about rewards in the environment [14] and in general also information (as defined in *information theory*) enters in this broader definition of rewards [15, 16].
Alongside the goal of maximizing reward it is important at this level to think about how the agent represents the task or, more precisely how it represents the environment with which it is interacting. Given the same task the goal could turn out to be very easy or very hard depending on the different representations chosen by the agent. Moreover representation of the environment as we have highlighted before in the example of the moth acts also as a principle to discriminate the biological plausibility of different algorithms.

Given the computational goal RL provides multiple algorithmic solutions to solve the problem. The correctness and efficacy of different RL algorithms depends critically on the quality of the *percept* highlighting again the centrality of the representation problem in this context.

Figure 1.1 shows a general classification of the different approaches in the space of *quality of the percept* vs. *knowledge of the environment* where

> **Quality of the percept** is related to how well the agent is able to represent the state of the environment
>
> **Knowledge of the environment** connects to how well the agent can predict

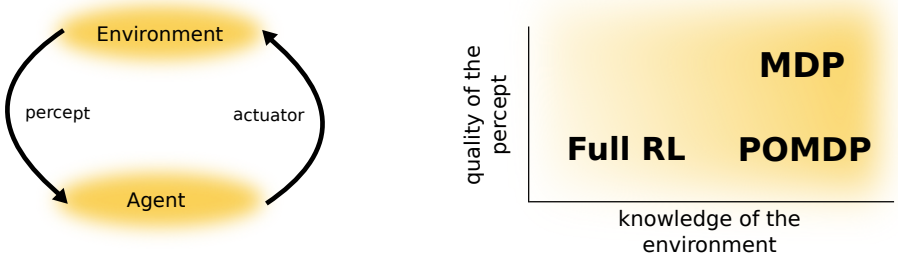


Figure 1.1: **Agent-environment interaction.** An agent capable to take actions interacts with the environment that returns a *percept* containing the information necessary to represent the state of the environment and the feedback (reward) on the agent's performance. The quality of the percept and how much the agent knows about the environment define a space of the algorithms in the reinforcement learning framework.

successive states upon taking a given action.

On the top-right corner there is the class of Markov Decision Processes (MDP) and the algorithms dedicated to this case go under the name of Dynamic Programming. In this case the agent has perfect knowledge of the states of the environment and knows also what is the effect of an action on the transition between states. This is a purely computational problem and does not require any form of learning.

If the agent receives poorer information from the environment (e.g. it just has a *belief* of which is the state of the environment) then we are in the bottom-right corner of Partially Observable Markov Decision Processes (POMDP). As we will see in the next chapter the cost to pay to keep the Markovian structure unaltered is that they POMDP must be solved in the continuous space of belief distributions. Algorithms facing this high complexity often look for approximate solution methods [17].

The last class, in the bottom-left corner, represents the full Reinforcement Learning problem. In this case the agent has a limited representation of the environment and does not know the effect of actions on the environment. For this reason it is forced to repeatedly interact with it. This aspect reveals the true nature of reinforcement learning and make two fundamental problems arise

Temporal credit assignment that is to determine which past actions deserve credit

Exploration-exploitation trade-off where the agent has to balance between *exploiting* what it already knows and *explore* in order to make better action selections in the future[2]

[2]The exploration-exploitation problem was already known in the engineering community studying optimal control under the name *dual control* problem (for an overview see [18])

The class of algorithms particularly suited to address this problems goes under the name of Temporal Difference algorithms (TD). As we will see in the next chapter TD-learning algorithms are specified by

- a **value function** saying how much a given action in a given state is valuable
- a **TD-error** essentially representing mismatches between outcomes and predictions,
- a **learning rate** determining the timescale over which reward histories are integrated to assess the value of an action
- a parameter defining the **time horizon** over which the behavior must be optimized.

Interestingly RL reaches also the implementational level, the third of Marr's diagram. Different experimental evidences [8, 19] pointed to the fact that dopaminergic neurons compute the TD-error and the learning rate is linked to serotonin, naturally related to what is called neural plasticity (see [20] and refs therein). Nowadays the mutual inspiration between the field of Reinforcement Learning and Neuroscience [21] shows that the separation between Marr's levels is just a matter of convenience and that they have necessarily to interact.

It is now clear how RL in its full generality permeates all the levels of Marr's paradigm and can be considered to be the appropriate framework to study behavior in a quantitative way.

All organisms have to face a complex environment and to do it they have to select meaningful environmental signals transforming them into actions. Learning and memory are two key features of animal adaptation to challenging situations. Wisely using information sensed from the environment and properly taking into account past experiences is often critical for optimal decision-making in a fluctuating environment and is involved in every aspect of an animal's life, from single individual tasks [22] to interaction with other individuals [23, 24].

Foraging and searching strategies are appealing problems to approach using a decision making framework. Both are critical activities for the survival and this aspect is thought to have profoundly shaped the organisms decision-making systems towards optimality. The ability to transform signals coming from the environment into actions is an aspect that crucially appears regardless the complexity of organisms. Learning and memory and how to deal with the perception-action cycle usually have been linked to cognitive capabilities of insects and vertebrates equipped with a central nervous system [25]. However even prokaryotes and eukaryotes show very well developed strategies. For example, simpler organisms such as bacteria sometimes cannot measure gradients in space. They present highly conserved signal transduction pathways to transform the temporal variability of chemical stimuli to precise chemotactic strategies corresponding to clockwise or counterclockwise rotation of

the flagellar motor. The goal in this case is to move towards higher concentration of chemoattractant or to excape from toxic substances [26]. Eucaryotes such as *Dictyostelium* are capable of measuring spatial gradients resulting into the activation of very well defined internal chemical pathways triggering locomotion with the consequent formation of pseudopods in the direction of the chemical gradient to reach the source [27]. Moreover, habituation considered a proof for the presence of learning mechanisms, has been identified in single cell organism such as *Physarum polycephalum* [28].

Given the ubiquity of behaviors classifiable as foraging or search strategies across organisms with radically different capabilities and the importance of these processes for species survival, it is likely that a range of different mechanisms from very simple rules to more complex ones have been selected under evolutionary pressure to cope with complex and unsteady environments.
We think that the framework we exposed presents the right characteristics to approach several problems related to behavior. It is general enough to be versatile but at the same time its mathematical formulation and the centrality of the algorithmic approach allows to frame biological phenomena inside a firm theoretical foundation.

Where can this thesis be placed in the described scenario ?

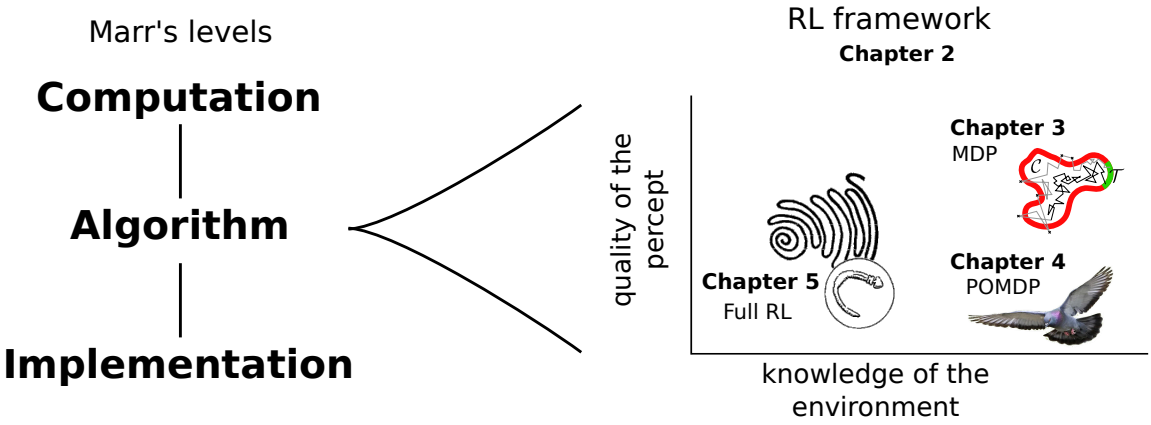


Figure 1.2: **Thesis organization.** Chapter 1 describes the general framework in which this Thesis can be placed. Chapter 2 presents the mathematical framework that will be used in Chapter 3 to describe a constrained search process by means of Markov decision processes (MDP), in Chapter 4 to study a problem inspired by olfactory navigation in pigeons using partially obervable Markov decision processes (POMDP) and Chapter 5 to show how optimal foraging strategies emerges in the context of full reinforcement learning.

The structure of the thesis is illustrated in Fig.1.2. The phenomena we are going to formalize in the framework of decision making theory are inspired by different

biological systems that we briefly present below. In the Marr's diagram this work can be placed at the algorithmic level. We will proceed from the top-right corner to the bottom-left one given in the RL classification scheme characterizing each biological phenomenon by means of a particular class of algorithms.

Constrained search processes

In Chapter 3 we will start studying a search process inside a confining domain. Diffusion and jump processes have been applied countless times to describe processes ranging from ecology to finance. In many instances the process is subject to a given number of constraints. Imagine for example to have an agent undergoing a random walk with the goal of reaching a target. What happens if we ask it to reach that target being confined within a given space region ? This situation can be described thinking about the agent to pay a cost each time it jumps outside the domain and to pay no cost otherwise. We show how the problem can be mapped into a particular class of Markov decision processes. As a particular example we consider a jump process with exponentially distributed jumps that have to reach a target without leaving a cylindrical channel. In this case we can analytically solve the Markov decision process and find the optimal policy corresponding to the transition probability of the jump process inside the cylinder. In analogy with the physics of confined polymer we investigate several geometrical properties of the constrained walks. Having the analytical expression for the transition probability of the constrained process we can efficiently characterize these properties in different regimes, going from the diffusive case to the one of long jumps corresponding to two different confinement regimes. The results presented in this Chapter are contained in

> Adorisio, M., Pezzotta, A., de Mulatier, C., Micheletti, C., Celani, A. (2018). Exact and Efficient Sampling of Conditioned Walks. *Journal of Statistical Physics*, 170(1), 79-100.

Navigating a noisy odor environment

Chapter 4 will be dedicated to approach from the algorithmic point of view the phenomenon of olfactory navigation in homing pigeons. The scientific community agrees on the fact that their ability to find home from unfamiliar places depend on their olfactory system.
In a map and compass mechanism olfaction gives the positional information and the compass is represented either by the coupling of the pigeon's internal clock with the position of the sun or by the geomagnetic field. In a simplified situation but considering anyway the experimental evidences accumulated in the last decades we cast this problem into a partially observable Markov decision process with Gaussian beliefs for which we can give a closed form solution for the optimal policy. The agent in this case is able to sense scalar signals representing the odors and use them

to optimally select actions. At variance from what has been proposed in the past, preliminary results show how this algorithm can explain experimental findings.

Optimal foraging strategies

Chapter 5 relates to the bottom-left corner in the algorithmic classification. The analysis of the trace fossils highlighted that the foraging behavior of different species evolved from very inefficient strategies showing self-intersecting trajectories to very compact and self-avoiding trails often showing stereotypical patterns such as spirals or meanders. In the field of ichnology, a sub-field of paleontology, a question that still does not have an answer is how this kind of strategies emerged during evolution. Previous models were able to reproduce these types of strategies using hard-coded instructions. Approaching the problem in the framework of reinforcement learning we show how the spirals and the meanders emerge as optimal solutions of a foraging problem. We will consider agents with very limited sensitivity able to represent the food distribution at very short spatial scales and with very simple locomotion capabilities much alike the ones that organisms of which we observe the trace fossils had. We will also investigate the task of searching for a target and eventually couple foraging and searching in a simple example of food exploitation.

Other projects

During these years I had the opportunity to collaborate on two other projects that are not presented in this Thesis. They both relate to chemotaxis approaching the phenomenon at two different scales. One is related to the flagellar motor and the other to the collective aspect of chemotaxis.

Flagellar motor: a cooperative binding model– *E.coli* is one of the model organisms to study bacterial chemotaxis. Its flagella are governed by rotary motors and their clockwise and counterclockwise rotation regulate the well known *run* and *tumble* behavior with which *E.coli* moves towards more favorable environments. In this work, exploiting the separation of time scales between different chemical processes regulating the flagellar motor, it is shown how to reduce the conformational spread model to a coarse-grained, cooperative binding model. This simplified model reproduces very well the switching dynamics between the clockwise and counterclockwise state. The results are contained in

> Pezzotta, A., Adorisio, M., Celani, A. (2017). From conformational spread to allosteric and cooperative models of E.coli flagellar motor. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(2), 023402.

Cooperative games and chemotaxis– Individuals in a group have often to cooperate to achieve common goals. We can think for example to multiple agents that

want to reach a given target minimizing a given cost function taking into account single individual costs and costs related to the interaction between the individuals. In this work it is shown that the equations that characterize the optimal strategy are identical to a long-known phenomenological model of chemotaxis. This allows to establish a dictionary that maps notions from decision-making theory to biophysical observables in chemotaxis, and vice versa. The results are contained in

> Pezzotta, A., Adorisio, M., Celani, A. (2018). Chemotaxis emerges as the optimal solution to cooperative search games. *Physical Review E*, in press.

My interest in biological systems started looking at multiple species interactions and their effect on the stability of ecosystems [29]. In these years I shifted my attention to the level of single agent behavior. It would be very interesting to see these two approaches converging at one point in the future to see how behavior and macroecological patterns are related.

Bibliography

[1] Marr, D. and Poggio, T. From understanding computation to understanding neural circuitry. 1976.

[2] Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[3] Cardé, R. T. and Willis, M. A. Navigational strategies used by insects to find distant, wind-borne sources of odor. *Journal of Chemical Ecology*, 34(7):854–866, Jul 2008.

[4] Celani, A., Villermaux, E., and Vergassola, M. Odor landscapes in turbulent environments. *Physical Review X*, 4(4), 2014.

[5] Vergassola, M., Villermaux, E., and Shraiman, B. I. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, 2007.

[6] Bau, J. and Cardé, R. T. Modeling optimal strategies for finding a resource-linked, windborne odor plume: Theories, robotics, and biomimetic lessons from flying insects. *Integrative and Comparative Biology*, 55(3):461–477, 2015.

[7] Martinez, D. and Martin-Moraud, E. Reactive and cognitive search strategies for olfactory robots. *Neuromorphic Olfaction*, 5:153–172, 2013.

[8] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition An Introduction*. 2018.

[9] Bouton, M. E. *Learning and Behavior: A Contemporary Synthesis*. Oxford University Press, 2016.

[10] Minsky, M. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1): 8–30, 1961.

[11] Bellman, R. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[12] Bellman, R. A markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

[13] Watkins, C. J. C. H. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

[14] Baldassarre, G. and Mirolli, M., editors. *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Berlin Heidelberg, 2013.

[15] Tishby, N. and Polani, D. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

[16] Bromberg-Martin, E. S. and Hikosaka, O. Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards. *Neuron*, 63(1): 119–126, 2009.

[17] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[18] Wittenmark, B. Adaptive dual control methods: An overview. *IFAC Proceedings Volumes*, 28(13):67–72, jun 1995.

[19] Schultz, W. Reward signals. *Scholarpedia*, 2(6):2184, 2007.

[20] Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F., and Dayan, P. An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, 9(1):10–12, 2018.

[21] Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258, jul 2017.

[22] Knaden, M. and Graham, P. The sensory ecology of ant navigation: From natural environments to neural mechanisms. *Annual Review of Entomology*, 61 (1):63–76, mar 2016.

[23] Ostrowski, E. A., Shen, Y., Tian, X., Sucgang, R., Jiang, H., Qu, J., Katoh-Kurasawa, M., Brock, D. A., Dinh, C., Lara-Garduno, F., Lee, S. L., Kovar, C. L., Dinh, H. H., Korchina, V., Jackson, L., Patil, S., Han, Y., Chaboub, L., Shaulsky, G., Muzny, D. M., Worley, K. C., Gibbs, R. A., Richards, S., Kuspa, A., Strassmann, J. E., and Queller, D. C. Genomic signatures of cooperation and conflict in the social amoeba. *Current Biology*, 25(12):1661–1665, jun 2015.

[24] Leadbeater, E. and Chittka, L. Social learning in insects — from miniature brains to consensus building. *Current Biology*, 17(16):R703–R713, aug 2007.

[25] Webb, B. Cognition in insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603):2715–2722, aug 2012.

[26] Wadhams, G. H. and Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nature Reviews Molecular Cell Biology*, 5(12):1024, 2004.

[27] Swaney, K. F., Huang, C.-H., and Devreotes, P. N. Eukaryotic chemotaxis: A network of signaling pathways controls motility, directional sensing, and polarity. *Annual Review of Biophysics*, 39(1):265–289, 2010.

[28] Boisseau, R. P., Vogel, D., and Dussutour, A. Habituation in non-neural organisms: evidence from slime moulds. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1829), 2016.

[29] Grilli, J., Adorisio, M., Suweis, S., Barabás, G., Banavar, J. R., Allesina, S., and Maritan, A. Feasibility and coexistence of large ecological communities. *Nature Communications*, 8, feb 2017.

Chapter 2

Methods

A look into optimal decision making theory

In this Chapter we focus on explaining the fundamental aspects of the optimal decision theory framework that especially deals with *sequential* decision-making problems. As explained in the introduction, Reinforcement Learning [1] and its formulation in terms of Markov decision processes[2] is the appropriate framework to study decision making and the optimality of behavior.

2.1 Markov Decision Processes

Markov Decision Processes (MDPs) represent a framework to formalize sequential decision-making and in general learning problems in stochastic domains. In this approach, an environment is modeled as a set of states and sequential actions can be performed by an agent to control the system's state. The agent has the goal to *act optimally* meaning that it has to control the system in such a way that some performance criterion is maximized.

More specifically we can imagine that the agent and the environment interact in a sequential way. At each time step t the agent receives some representation of the environment's *state*, $s_t \in \mathcal{S}$ and based on this it selects an action $a_t \in \mathcal{A}$ receiving a *reward* r_{t+1} and eventually finding itself in a new state, s_{t+1}[1]. Thus the interaction between the agent and the environment can be summarized by a *trajectory* of the form

$$s_0, a_0, r_1, s_1, a_1, r_2, s_2, \dots s_t, a_t, r_{t+1}, s_{t+1}$$

We can define the transition probability $p(s', r|s, a)$ as the probability of arriving in state s' after taking action a in state s and receiving a reward r. It represents a

[1] we write r_{t+1} instead of r_t to emphasize that action a_t jointly determines the reward and the new state r_{t+1}, s_{t+1}.

Figure 2.1: **Perception-action cycle in the MDP.** The agent receives the percept that is composed by a reward and the true state of the environment. According to the state the agent selects an action following a given policy.

central object in MDP theory and is sometimes called the *model of the environment*. For $p(s', r|s, a)$ to be a proper probability distribution over possible next states s' and rewards r we require that

$$\sum_{s',r} p(s', r|s, a) = 1 \ \forall s, a$$

The transition probability $p(s', r|s, a)$ completely characterize the environment's dynamics under the agent's actions. We are assuming implicitly that the transition between states are Markovian. Considering the history of the interactions between agent and environment this means

$$p(s_{t+1}, r_{t+1}|s_0, a_o, s_1, a_1, s_2, \ldots s_t, a_t) = p(s_{t+1}, r_{t+1}|s_t, a_t)$$

highlighting that the state s_t carries enough information to make an optimal decision in the future.

Informally we can state that the *goal* of the agent is to maximize[2] the expected value of the sum of rewards. To be more specific the agent seeks to maximize the expected value of *return* defined by

$$R_t = r_{t+1} + r_{t+2} + \ldots r_T \quad \text{[return]} \tag{2.1}$$

where T is the final step. This return definition makes sense in the case of *episodic tasks* (i.e. when the agent reaches a *terminal state* and the interaction with the environment restarts).
There is also the case of *continuing tasks* in which we cannot identify an episode and

[2]We use reward to define a *scalar* signal received by the agent from the environment. Obviously it can be positive or negative (a *cost* paid by the agent) and in this case the goal will be to minimize the cost.

the agent-environment interaction continues until $T = \infty$. The expected return (2.1) will consist of a sum of infinite terms and could turn out to be infinite itself. To avoid this situation we introduce the *expected discounted return*

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad \text{[discounted return]}. \tag{2.2}$$

The parameter γ satisfying $0 \leq \gamma \leq 1$ characterizes the "nature" of the agent. A $\gamma = 0$ agent wants to maximize the immediate reward (i.e. $R_t = r_{t+1}$). As γ approaches 1 the agent weights future rewards more strongly. We can unify the two expressions for the return (eq.(2.1) and (2.2)) admitting that an episodic task reaches an absorbing state s_A that transitions only to itself generating zero reward (see Fig. 2.2).

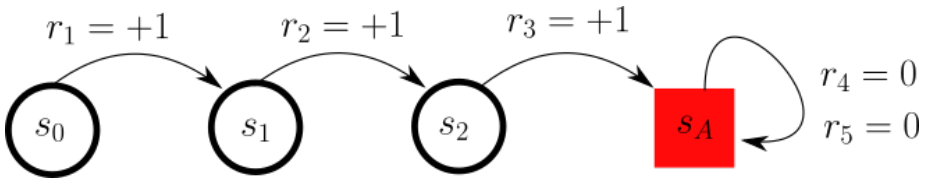


Figure 2.2: Starting from s_0 we accumulate $R_0 = r_1 + r_2 + r_3 + 0 + 0 + 0 + \ldots$ obtaining the same reward whether we sum over the first $T = 3$ steps or over the infinite sequence.

Summarizing the return can be expressed by

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1} \tag{2.3}$$

where $T = \infty$ or $\gamma = 1$ but not both conditions together.
A nice property of the return is that it satisfies

$$R_t = r_{t+1} + \gamma R_{t+1} \tag{2.4}$$

This property is at the base of what we are going to present in the following section.

2.2 Value function: how valuable is a policy

The agent interacts with the environment following a given policy. Formally a policy $\pi(a|s)$ is a mapping from states to probabilities of selecting each possible action. Following the policy π means that at time t the agent has the probability $\pi(a|s)$ of taking action $a_t = a$ if in state $s_t = s$. Thus it would be useful for the agent to have a "measure" of how good a given state is or how good is taking an action in a given state. In other words we would like to give a value to each state given the policy π.

We can define the *state-value function* of a state s under a policy π as the expected return when starting in $S_t = s$ and following π thereafter. Formally we can define it as the average value according to π of the return starting from $s_t = s$

$$V_\pi(s) \equiv \mathbb{E}_\pi \left[R_t | s_t = s \right] \tag{2.5}$$

with the definition of R_t given in eq. (2.3). Similary if the agent wants to know how good it is to take action $a_t = a$ when in $s_t = s$ we define the *action-value function*

$$Q_\pi(s, a) \equiv \mathbb{E}_\pi \left[R_t | s_t = s, \ a_t = a \right] \tag{2.6}$$

Using eq. (2.4) to simplify eq. (2.5), we can write

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi \left[r_{t+1} + \gamma R_{t+1} | s_t = s \right] \\ &= \sum_{a, s', r} p(s', r | s, a) \pi(a|s) \left[r + \gamma V_\pi(s') \right] \end{aligned} \tag{2.7}$$

For the action-value function (2.6) reasoning the same way we obtain

$$Q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma \underbrace{\sum_{a'} \pi(a'|s') Q_\pi(s', a')}_{V_\pi(s')}] \tag{2.8}$$

The last two equations represent two important recursion relations for the value-function and action-value function respectively.[3] Solving a MDP means finding an *optimal policy* that is the policy that roughly speaking accumulates the largest amount of reward for the prescribed task. Thus given a set of policies we look for the optimal one π^* that will have $V_{\pi^*}(s) = V^*(s) = \max_\pi V_\pi(s) \ \forall s$. Accordingly $Q^*(s, a) = \max_\pi Q(s, a) \ \forall s, a$

Concluding, if we identify $V^*(s) = \max_a Q^*(s, a)$[4], the optimal policy is such that the following Bellman optimality equations for V^* and Q^* hold

$$\begin{aligned} V^*(s) &= \max_a \sum_{s', r} p(s', r | s, a) \left[r + \gamma V^*(s') \right] \\ Q^*(s, a) &= \sum_{s', r} p(s', r | s, a) [r + \gamma \max_{a'} Q^*(s', a')] \end{aligned} \tag{2.9}$$

The solution for the MDP (i.e. the optimal policy) will be given by

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{2.10}$$

[3] we can define $T(s', a, s) \equiv p(s', r | s, a) \pi(a|s)$ as the transition probability from s to s' taking action a.

[4] Intuitively the optimal value for the state s when acting with the optimal policy is the total return we get choosing the best action from that state.

Notice that the structure of eq. (2.9) reflects the property (2.4)

$$V^* = \texttt{immediate reward} + \gamma(\texttt{expected value for the future})$$

Summarizing an MDP with optimal policy π^* is defined by

- set $\mathcal{S}$ of states
- set $\mathcal{A}$ of actions
- the model of the environment $p(s'|s,a)$ $\forall a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$ giving the probability to end in state s' when starting in s taking action a
- the return $R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$ with $0 \leq \gamma \leq 1$ ($\gamma = 1$ only for $T < \infty$)

$$\pi^* = \arg\max_a V^*$$

with V^* solution of

$$V^*(s) = \max_a \sum_{s',r} p(s', r|s, a) \, [r + \gamma V^*(s')(]$$

We introduced the structure of the MDP and the Bellman equation that must be solved in order to find the optimal policy. There are many methods to solve an MDP and all these techniques go under the name of Dynamic Programming. Policy iteration is one of them and the idea is summarized in Fig. 2.3. Given a policy we *evaluate* it finding its value V. We then *improve* the previous policy selecting greedy actions on the computed value and evaluate again the new policy. If both the *evaluation* process and the *improvement* process stabilize, that is, no longer produce changes, then the value function and policy must be optimal.

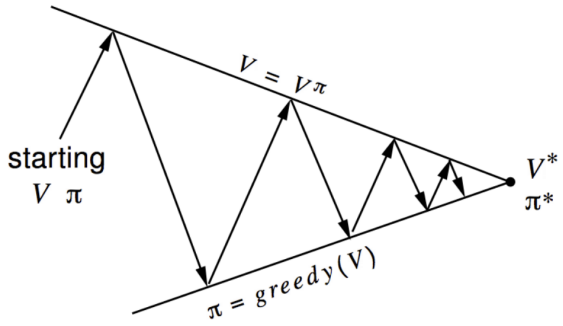


Figure 2.3: **Policy iteration scheme.**

In Chapter 3 we are going to present how to condition a jump process to stay inside a given domain. We will highlight the equivalence of this problem with a particular class of Markov decision processes and solve for the optimal policy.

2.3 Partially observable Markov decision processes

If an agent has access to the true state of the environment and it knows the model then the problem can be described as a Markov decision process. What if the agent lack precise knowledge of state knowing only the probability to be in that state ? In this case the MPD framework can be extended to take into account that the agent does not have access to the real state s of the environment but only to observables related to it. The framework of Partially Observable Markov Decision Processes (POMDPs) allows for principled decision-making under uncertainty in the ability of representing the state of environment. In this section we present the POMDP structure in relation to what we already introduced for MDPs. Even if we do not focus in details on the algorithms solving POMDPs we discuss the ideas that are central to understand the results of Chapter 4.

The first property to relax when defining a POMDP in relation to the MDP framework is the assumption that the agent knows with full certainty the state of the environment. The agent interface with the environment is represented by sensors that are capable of extracting few observables among all the ones that define the true state of the environment. The Markov property of the (fully observable) MDP is based on the fact that the agent has access to the true state and that the true state is *sufficient* to represent the environment. In general, when certain state features are hidden from the state signal received by the agent the assumption about Markovianity ceases to hold. Partial observability in general arises because the sensors of an agent represent different environmental states in the same way and also because the sensors return noisy measures. The main result of what we present below is that we can reacquire the Markov property disrupted by partial osbervability by paying the cost of formulating a MDP in the continuous space of probabilities over the states (the *belief* state) [5].

To define a POMDP we can build on top of the definition of a MDP given in the box at the end of section 2.1. In the sequential interaction with the environment the agent takes an action a when in state s and the environment transitions to state s' according to $p(s'|s,a)$ and the agent receives a reward r. The big difference with the MDP setting is that the agent now perceives an observation o related to s' instead of observing s' itself. As we said before the observation is in general a noisy quantity. We can encode this fact saying that there is a *likelihood* $\ell(o|s)$ to obtain the observation o when in state s [6]. The problem of dealing with observations is that a direct mapping of observations to action is not sufficient to act optimally and a sort of memory of the past experience is needed. A first idea could be to book-keep the history of observations received and actions taken. Even if it sounds intuitive

[5] The Bellman equation presented in the previous section will be defined on the space of probabilities over states and not anymore on the set of states

[6] The likelihood is normalized $\sum_o \ell(o|s) = 1 \ \forall \ s$

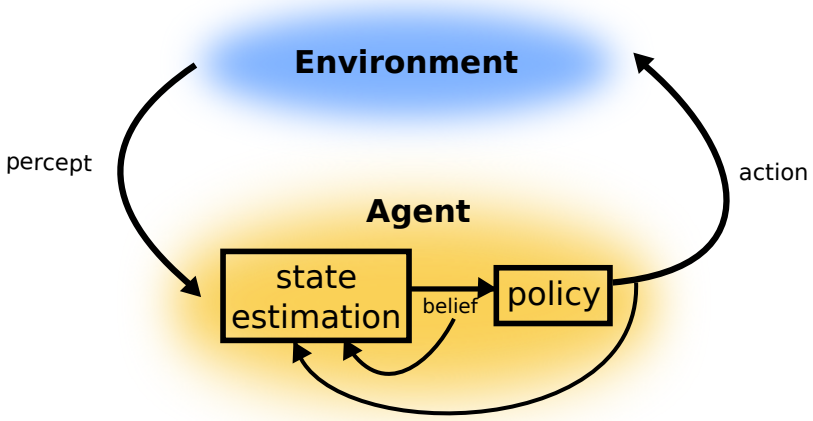


Figure 2.4: **The perception-action cycle in the POMDP.** The agent receives from the environment the percept that contains the reward and an observation. The agent exploits the observation to update the belief and selects an action according to the new generated belief.

it is not the most efficient way to do it especially if we consider continuing tasks. Fortunately there is a better option that is the one to transform all the information about the past into a *belief* over the states. As the word says the belief distribution $b(s)$ is the probability to be in state s giving an idea about the "belief" the agent has to be in state s. Obviously we must require $\sum_s b(s) = 1$. To compress the past history of actions and observations into the belief over the states we need both the model of the environment $p(s'|s,a)$ and the likelihood $\ell(o|s)$. We look for a rule that allows us to update the belief as soon as an observation is received by the agent. This rule must compute a new belief state $b'(s')$ given the old belief $b(s)$, observation o and action a. A natural choice is to use a *Bayesian update rule* for the belief $b(s)$. In fact, making the sequential time dependence explicit, according to Bayes rule we have that

$$\begin{aligned} b_{t+1}(s_{t+1}|a_t) &\equiv Prob(s_{t+1}|b_t, a_t, o_{t+1}) \\ &= Prob(o_{t+1}|s_{t+1}, a_t) Prob(s_{t+1}|b_t, a_t) \\ &= \frac{\ell(o_{t+1}|s_{t+1}, a_t) \sum_{s_t} p(s_{t+1}|s_t, a_t) b_t(s_t)}{p(o_{t+1}|a_t, b_t)} \end{aligned} \tag{2.11}$$

where $p(o_{t+1}|a_t, b_t) = \sum_{s_{t+1}} \ell(o_{t+1}|s_{t+1}, a_t) \sum_{s_t} p(s_{t+1}|s_t, a_t) b_t(s_t)$ [7]. If in MDPs we have that a policy maps states into actions in POMDPs the policy maps a belief into

[7] $p(o|b,a)$ it's a normalization of the belief b_{t+1} but it can be interpreted as the transition probability between b_t and b_{t+1} when the agent takes action a_t (see footnote 8)

actions. As in MDPs the goal of the agent is to maximize the discounted reward. Finding the optimal policy requires to give a value to a given policy starting from a given belief and in analogy to (2.5) we can define the value function for POMDPs as

$$V_\pi(b) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R(b_t) \mid b_0 = b \right] \tag{2.12}$$

where $R(b_t) = \sum_s b_t(s) r(s,a)$ and b_0 is the initial belief. Concluding, it can be shown that an optimal policy for a POMDP is the solution of the optimal Bellman equation that in this case takes the form [8]

$$V^*(b) = \max_a \left[\sum_s r(s,a) b(s) + \gamma \sum_o p(o|b,a) V^*(b) \right] \tag{2.13}$$

and the optimal policy is

$$\pi^*(b) = \arg\max_a \left[\sum_s r(s,a) b(s) + \gamma \sum_o p(o|b,a) V^*(b) \right] \tag{2.14}$$

As in the MDP case the equation for the value keeps the structure

$$V^* = \texttt{immediate reward} + \gamma(\texttt{expected value for the future})$$

Concluding, we have summarized how a POMDP is equivalent to an MDP in the continuous belief space. The following box, to be compared with the one at the end of Sec. 2.2, brings together all the ingredients that define a POMDP.

[8] In analogy to the MDP case where we defined $T(s', a, s)$ (see footnote 3) we can interpret the term $p(o|b,a)$ as the transition probability from b to b' taking action a if we rewrite $T(b', a, b) \equiv \sum_o \mathcal{I}(b' - b|o,a) p(o|a,b)$ where

$$\mathcal{I}(b' - b|o,a) = \begin{cases} 1 & \text{if } b' \text{ and } b \text{ are related by (2.11)} \\ 0 & \text{otherwise} \end{cases}$$

- set $\mathcal{S}$ of states
- set $\mathcal{A}$ of actions
- the model of the environment $p(s'|s,a)$ $\forall a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$ and a likelihood function $\ell(o|s')$ to encode observations into the belief distribution over states $b(s)$
- a rule to update belief $b(s)$ according to observations
- the return $R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$ with $0 \leq \gamma \leq 1$ ($\gamma = 1$ only for $T < \infty$)

$$\pi^*(b) = \arg\max_a V^*(b)$$

with V^* solution of

$$V^*(b) = \max_a \left[\sum_s r(s,a)b(s) + \gamma \sum_o p(o|b,a)V^*(b)\right]$$

2.4 Full Reinforcement Learning

The aim of Reinforcement Learning (RL) algorithms is to find an optimal policy that maximizes the return without relying on the model of the environment. Because the model of the MDP is not known the agent has to *interact* with the environment and *explore* it receiving scalar feedback in the form of rewards. In this situation the agent has to deal with a trade-off between *exploring* newer actions that could possibly return higher rewards and *exploiting* what it already knows.

In the generalized policy-iteration approach one crucial step is the *policy evaluation* or *prediction* step. Given a policy we must compute its value

$$V_\pi(s) = \mathbb{E}_\pi \left[R_t | s_t = s\right] \tag{2.15}$$
$$= \mathbb{E}_\pi \left[r_t + \gamma V_\pi(s_{t+1}) | s_t = s\right] \tag{2.16}$$

As in eq. (2.15) we can estimate V_π sampling R_t and then adjusting the estimate of the value as

$$V(s_t) \leftarrow V(s_t) + \alpha \left[R_t - V(s_t)\right]$$

where α is called learning rate.
This kind of algorithm goes under the name of Monte Carlo Reinforcement Learning. Without requiring a model of the environment they approximate the value function because they sample the reward at each time step computing the return R_t only a the end of the episode. Apart from being very intuitive this approach could become infeasible in the case of very long episodes or for continuing tasks.

On the other side there are Dynamic Programming (DP) approaches that evaluate the policy approximating the value function as in eq. (2.16). Even if in the DP approach the model of the environment is known, the value function is approximated because it is computed using the current estimate $V_t(s_{t+1})$ instead of the unknown $V_\pi(s_{t+1})$ until the process converges. Using estimates of a quantity to compute other estimates the quantity is called *bootstrapping*.

There is a third class of algorithms that goes under the name of Temporal Difference (TD) algorithms [3] combining both approximation aspect of *sampling* and *bootstrapping* that we are going to briefly discuss in the following.
In general these algorithms have the form

$$V_{t+1}(s) \leftarrow V_t(s) + \alpha_t \delta_t(\gamma, V_t).$$

where α_t is again the learning rate and the quantity δ_t is called the TD error. We made the dependence explicit on γ and the current value function V_t because what differentiates the algorithms in RL is the way the TD-error is computed. In general the idea is to write the TD error as

$$\delta_t = \texttt{target} - \texttt{current value}.$$

We will highlight this structure when presenting two RL algorithms below. In the following we present Q-learning and SARSA algorithms. The Bellman equation for the action-value function (eq. 2.9) represents the perfect starting point to understand their structure.

Q-learning

Q-learning [4] is one of the most basic and popular way to estimate the action-value function Q and it takes the form

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t \left[\underbrace{r + \gamma \max_a Q_t(s_{t+1}, a)}_{\texttt{target}} - Q_t(s_t, a_t) \right] \tag{2.17}$$

This update rule can be understood rewriting the optimal Bellman equation 2.9 as

$$\mathbb{E}_{\pi^*}[r + \gamma \max_{a'} Q^*(s', a') - Q^*(s, a)] = 0$$

telling us that at optimality the expectation of the TD-error satisfies

$$\delta^* = r + \gamma \max_a Q^*(s_t, a) - Q^*(s, a) = 0.$$

Since we do not know the model of the environment we must substitute $\mathbb{E}_{\pi^*}[\cdot]$ with a sampling step where we generate an experience $(s_{t+1}, a_t, r_{t+1}, s_t)$ interacting with the environment following policy π.
Q-learning is called an *off-policy* algorithm in the sense that even if we keep exploring actions to balance exploration and exploitation following a given policy π it is actually evaluating the optimal policy π^*. The convergence to the optimal Q is guaranteed under the assumption that each state-action pair is visited infinitely many times and the learning rate α_t is decreased appropriately [9].

SARSA

If Q-learning is an off-policy algorithm SARSA [3] represents an *on-policy* approach with the following update rule

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t \left[\underbrace{r + \gamma Q_t(s_{t+1}, a_{t+1})}_{\texttt{target}} - Q_t(s_t, a_t) \right] \tag{2.18}$$

Starting from the Bellman equation (2.8) for Q we can rearrange it as

$$\sum_{s',r} p(s', r|s, a)[r + \gamma \sum_{a'} \pi(a'|s') Q_\pi(s', a') - Q_\pi(s, a)] = 0.$$

The quantity inside the square brackets give us a hint on how to interpret the TD-error for SARSA algorithm. As before, we do not know the transition probabilities and we approximate the expectation value with a sample of $(s_{t+1}, a_{t+1}, r, s_t, a_t)$.[10].
At variance from what happens in Q-learning, SARSA uses the policy being evaluated to extract a_{t+1}. This is why it is considered an on-policy algorithm. Convergence to the optimal Q [5] is guaranteed also in this case by the usual stochastic approximation conditions (see footnote 9).

Neurons and TD-error computation

The interesting point about RL algorithms is that it has been shown that dopaminergic neurons [1, 7] encode the TD-error and a positive or negative δ is related to

[9] Watkins and Dayan (1992) prove the convergence of Q-learning algorithm. In general convergence analysis of TD learning algorithm is rooted in stochastic approximation analysis. Q-learning and SARSA [5] algorithm (see next section) converge with probability 1 to the value function if

$$\sum_t \alpha_t = \infty \quad \sum_t \alpha_t^2 < \infty$$

or in the mean if α is kept fixed and sufficiently small. There is a wider class of algorithm called TD(λ) that has been proven to converge with probability 1 [6]. λ is a parameter related to how far in the future the agent can look (MC algorithms corresponds to TD(1) and SARSA to TD(0).)

[10] the name SARSA comes from the structure of the experience: s', a', r, s, a

a "good" or "bad" action because it led to a state with a better or worse than expected value. Moreover there is evidence (see [8] and references therein) that other neuromodulators different from dopamine are responsible for the representation of the other RL parameters such as the discounting factor γ and the learning parameter α. In particular the learning rate α has been hypothesized to be connected to neural plasticity regulated by serotonin, another neuromodulator [9].

Bibliography

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition An Introduction*. 2018.

[2] Bellman, R. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[3] Sutton, R. S. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 1988.

[4] Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 1992.

[5] Singh, S., Jaakkola, T., Littman, M. L., and Szepesvári, C. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine Learning*, 38(3):287–308, 2000.

[6] Dayan, P. and Sejnowski, T. J. TD(λ) Converges with Probability 1. *Machine Learning*, 1994.

[7] Schultz, W. Reward signals. *Scholarpedia*, 2(6):2184, 2007.

[8] Doya, K. Modulators of decision making. *Nature Neuroscience*, 11(4):410–416, 2008.

[9] Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F., and Dayan, P. An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, 9(1):10–12, 2018.

Chapter 3

Optimal control of jump processes

Freely jointed chains confined in a channel

1 Introduction

Random walks are ubiquitous in physics and have countless applications in biological systems [1–3], ecology [4], finance [5], chemistry and transport phenomena [6]. In many instances, the walk is subject to a certain number of global constraints, e.g. the walker can be restricted to a given domain of space or constrained to reach a certain target.

Generating constrained walks by stochastic techniques is computationally challenging [7]. This is readily illustrated by considering the inherent inefficiency of simple resampling strategies. In such approaches one could, in principle, generate by Monte Carlo or other schemes, a large ensemble of unrestricted walks and then reject *a posteriori* those violating the constraints. This naive strategy is bound to incur in a rejection rate that increases exponentially fast with the walk length.

To overcome these challenges, various advanced stochastic sampling methods have been proposed (see e.g. [8]). These techniques work by suitably biasing the system towards the relevant configurations that would otherwise be visited only exceptionally in the unconstrained case. In practice, finding an algorithm able to generate efficiently constrained random walks with the exact statistical weight still remains a daunting task.

A different approach was pioneered by Doob [9] for diffusive processes and recently revisited in the physics literature, see e.g. [10–13]. In short, this method is based on the observation that any constrained random walk is exactly equivalent to an auxiliary unconstrained one, via a suitable reweighting of the transition probability. Clearly, the unconstrained version of the original process is typically much more amenable than the former to computational, and even analytical treatment. The challenge, in this case, lies in how to exactly derive the auxiliary process from the given global constraints.

Another interesting perspective is given by looking at the problem from the point of view of decision processes. In particular in this Chapter, we show how the unconstrained auxiliary process can be obtained both by reweighting and as the solution of a particular class of Markov decision processes [14–17].

We illustrate our approach with the example of a walk confined inside a cylinder and forced to reach one of its ends. For this problem the exact expression of the transition probability for the jumps can be obtained analytically, providing an efficient way to generate constrained walks.

This method allows to easily generate constrained jump processes with an arbitrarily number of jumps. The numerical and analytical control we have on this specific problem allowed us to study different geometrical properties of the walks. For example we were able to study the transition between weak and strong confinement that would be otherwise difficult to access for asymptotically long chains. We also explore possible connections with the physics of confined polymers, still a very active field of research [18–20].

2 Constrained random walks

Consider a random walk in a domain $\mathbb{R}^d$ where the sequence of visited points $\{x_i\}_{i\geq 0}$ follows from a transition probability $p(x_{i+1}\,|\,x_i)$ for the jumps $x_i \to x_{i+1}$ Let us now choose a domain $\mathcal{C}$ inside which we want to constrain the trajectories. We also define a terminal domain $\mathcal{T}$, outside $\mathcal{C}$, where the last point of the trajectory has to land. Therefore, the ensemble of constrained trajectories we are considering corresponds to unbiased realizations of the jump process that, starting anywhere in $\mathcal{C}$, happen to stay inside this domain for all jumps except for the last step that takes them inside $\mathcal{T}$.

The joint probability density function of the trajectory $(x_1, \ldots, x_T)$ starting from $x_0 \in \mathcal{C}$ and reaching x_T at time T reads

$$P(x_1, \ldots, x_T \,|\, x_0) = \prod_{i=1}^{T} p(x_i \,|\, x_{i-1})\,. \tag{1}$$

Our goal is to find the subset of trajectories generated by P that stay inside $\mathcal{C}$ and terminate in $\mathcal{T}$. The new joint pdf of the trajectory $(x_1, \ldots, x_T)$ under the confinement condition can thus be written as

$$Q(x_1, \ldots, x_N \,|\, x_0) = \frac{\prod_{i=1}^{T} p(x_i \,|\, x_{i-1})\, \mathbb{I}_{\mathcal{C}}(x_i)}{Z_T(x_0)}\,, \tag{2}$$

where

$$\mathbb{I}_{\mathcal{C}}(x) = \begin{cases} 0, & x \notin \mathcal{C} \\ 1, & x \in \mathcal{C} \cup \mathcal{T} \end{cases}$$

The normalization factor

$$Z_T(x_0) = \int \mathrm{d}x_1 \dots \mathrm{d}x_T \prod_{i=1}^{T} p(x_i \,|\, x_{i-1})\, \mathbb{I}_{\mathcal{C}}(x_i)\,. \tag{3}$$

is the probability that a trajectory originating from x_0 sampled from P does not leave $\mathcal{C}$ and terminates in $\mathcal{T}$. Notice that this term can be rewritten in the following recursive form

$$Z(x_0) = \int \mathrm{d}x_1 p(x_1|x_0)\mathbb{I}_{\mathcal{C}}(x_1) \overbrace{\int \mathrm{d}x_2 \cdots \mathrm{d}x_T \prod_{i=2}^{T} p(x_i|x_{i-1})\mathbb{I}_{\mathcal{C}}(x_i)}^{Z(x_1)}\,. \tag{4}$$

Exploiting this backward equation we can rewrite eq.(2) as

$$Q(x_1 \dots x_T|x_0) = \frac{1}{Z(x_T)} \cdot \prod_{i=1}^{T} \frac{p(x_i|x_{i-1})\mathbb{I}_{\mathcal{C}}(x_i)Z(x_i)}{Z(x_{i-1})}. \tag{5}$$

with $Z(x_T) = 1$, meaning that the walker reached the terminal domain $\mathcal{T}$.

The key non trivial fact is that the constrained trajectories described by $Q(x_1 \dots x_N|x_0)$ still follow a Markov process described by the well defined[1] transition probability

$$q(x'|x) = \frac{Z(x')}{Z(x)}\, p(x'|x)\, \mathbb{I}_{\mathcal{C}}(x'). \tag{6}$$

Summarizing, the constrained Markovian trajectories inside $\mathcal{C}$ will be described by

$$q(x'|x) = \begin{cases} \frac{Z(x')}{Z(x)} p(x' \,|\, x) & \text{for } x' \text{ in } \mathcal{C} \text{ or } \mathcal{T}, \\ 0 & \text{elsewhere,} \end{cases} \tag{7}$$

where the weighting function $Z(x)$ obeys the *linear* integral equation

$$Z(x) = \int_{\mathcal{C}} \mathrm{d}x'\, Z(x')\, p(x' \,|\, x) \qquad \text{for all } x. \tag{8}$$

Note that Z is the probability of being absorbed in $\mathcal{T}$ before being absorbed elsewhere outside $\mathcal{C}$ (see [9–12]).

It is also important to remark that the constrained process (7) is not equivalent to enforcing reflecting boundary conditions at the frontier of $\mathcal{C}$.

[1] Equation (6) ensures that each $q(x_i \,|\, x_{i-1})$ is correctly normalized.

An interesting fact is that eq.(7) for the new transition probability and eq. (8) defining the reweighting factor assume a specific meaning in the framework of optimal controlled Markov decision processes. Using the ideas presented in Sec.2.1 in Chapter 2 we show in the next section how the transition probability $q(x'|x)$ corresponds to the optimal policy of a Markov decision process and that $Z(x)$ is strictly related to the value function.

3 Reweighting and Markov decision processes

Let us imagine that the jump process described in the previous section by the transition probability $p(x'|x)$ is the trajectory of an agent inside a given region of the space $\mathbb{R}^d$. A subregion $\mathcal{T}$ of the whole space represents the target that the agent wants to reach while remaining inside $\mathcal{C}$. The domain $\mathcal{T}$ represents an absorbing state and to force the agent to stay inside $\mathcal{C}$ we penalize it when it jumps outside $\mathcal{C} \cup \mathcal{T}$. The cost it has to pay is

$$c(x) = \begin{cases} k, & x \notin \mathcal{C} \cup \mathcal{T} \\ 0, & \text{otherwise} \end{cases}$$

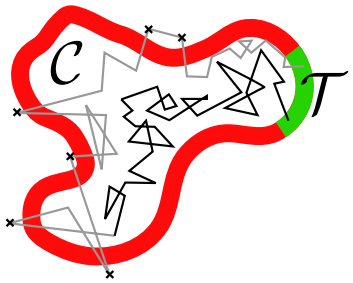


Figure 3.1: We would like to select only the trajectories that never leave the red domain and arrive at the green target. Each time a trajectories jumps out of the domain it accumulates a cost k.

Thus what the agent would like to do is to control its trajectory to minimize the cost. We give the agent the freedom to reshape its original dynamics described by $p(x'|x)$ defining a new transition probability $p_u(x'|x)$ to achieve the goal. To make a parallel with what we presented in Chapter 2 we can think to p_u as the analogous of the actions. There we defined the transition probability $p(x'|x,u)$ from state x to state x' taking action u chosen in a discrete set of actions. Here we can imagine that the agent has the power to directly choose the transition probability, i.e. $p(x'|x,u) \equiv p_u(x'|x)$. A path of this Markov process will be described by

$$\mathbb{P}_u(x_1, \ldots, x_N|x_0) = \prod_{i=1}^{T} p_u(x_i|x_{i-1}). \tag{9}$$

However the agent has not an infinite freedom on the choice of $p_u(x'|x)$ but it must pay for deviating from its original dynamics $p(x'|x)$. Intuitively this prevents the controlled process to jump directly to the target. Then the agent has to deal with two types of cost. One given by the constraint to stay inside $\mathcal{C}$. The other cost comes from the possibility to control its dynamics. To incorporate these two types of cost in a single cost function we can define

$$\mathcal{L} = \langle c(x') \rangle_{p_u} + D_{KL}(p_u||p) \tag{10}$$

where $D_{KL}(p_u||p)$ is the Kullback-Leibler divergence between the controlled and uncontrolled process defined by

$$D_{KL}(p_u||p) = \sum_{x'} p_u(x'|x) \log \frac{p_u(x'|x)}{p(x'|x)}.$$

In this context $D_{KL}(p_u||p)$ naturally emerges as a measure to quantify the cost for the control. Moreover it must be $p_u(x'|x) = 0$ whenever $p(x'|x) = 0$ to make the Kullback-Liebler divergence well defined and to keep the same forbidden transitions in the two Markov processes.

The Bellman equation for this problem is

$$V(x) = \min_{p_u} \Big[\mathcal{L} + \langle V(x') \rangle_{p_u} \Big] \tag{11}$$

where the usual value function $V(x)$ now is the optimal *cost-to-go* function, defined as the expected cumulative cost for starting at state x and acting optimally thereafter. Given the definition of the cost function (10) the Bellman equation can be rewritten as

$$V(x) = \min_{p_u} \Big[\sum_{x'} p_u(x'|x) \log \frac{p_u(x'|x) e^{c(x')+V(x')}}{p(x'|x)} \Big].$$

We can now define

$$\tilde{p}(x'|x) \equiv \frac{1}{Z(x)} e^{-c(x')-V(x')} p(x'|x) \tag{12}$$

where $Z(x)$ for the moment represents just the normalization factor

$$Z(x) = \sum_{x'} e^{-c(x')-V(x')} p(x'|x). \tag{13}$$

The Bellman equation now takes the form

$$V(x) = \min_{p_u} D_{KL}(p_u||\tilde{p}) - \log Z(x). \tag{14}$$

The Kullback-Leibler reaches its minimum if $p_u = \tilde{p}$ implying that

$$V(x) = -\log Z(x).$$

The normalization condition (13) and the controlled transition probability p_u now takes the form

$$\begin{aligned} p_u(x'|x) &= \frac{Z(x')}{Z(x)} e^{-c(x')} p(x'|x) \\ Z(x) &= \sum_{x'} Z(x') e^{-c(x')} p(x'|x). \end{aligned} \tag{15}$$

If the agent pays an infinite cost when it jumps outside the domain $\mathcal{C}$ (i.e. $k \to \infty$ in eq.(3)) then the equations above are exactly the same as eq. (7) and (8) showing the analogy between Markov Decision processes and the reweighting procedure presented in the previous section. This class of problems goes under the name of linearly solvable Markov decision processes [17]. As already pointed out at the beginning of the section in this case the notion of "action" defined in the general framework of MDP is replaced by the transition probability $p_u(x'|x)$, somehow generalizing the concept of policy as a mapping between states and actions.

The reference dynamics $p(x'|x)$ is arbitrary and must reflect the properties of the system under exam. For example in the next section inspired by polymer physics we are going to see how to constrain a jump process inside a cylindrical channel. In this case that can be solved analytically, $p(x'|x)$ will described a jump process with exponentially distributed jump length. Exploiting the analogy with the searcher in this case the goal will be to reach a target region at the end of the cylinder without leaving it.

4 Constraining a jump process inside a cylindrical channel

In general, the linear integral equation (8) can be solved numerically and once the weights are obtained, the deformed process q can be used to generate samples of the constrained ensemble with their exact statistical weight.

In this section we show an interesting case where the problem can be solved analytically providing a very effective method to sample configurations that would otherwise be exceedingly rare in an unbiased sampling. As we show below, one such instance is when the confining regions is a cylinder. Besides being amenable to extensive characterization within the aforementioned framework, this system was

chosen for its connection with confined polymer chains, a classic and yet still actively investigated topic in polymer physics.

Let us consider the following free jump process in $\mathbb{R}^3$

$$p(x_{i+1} \,|\, x_i) = \frac{m^2 \, e^{-m\,||x_{i+1}-x_i||}}{4\pi \,||x_{i+1} - x_i||} \,, \tag{16}$$

where m is a parameter controlling the mean length of a jump,

$$\ell_f \equiv \langle ||x_{i+1} - x_i|| \rangle_p = 2/m \,. \tag{17}$$

The transition probability of Eq. (16) can be obtained by considering a three-dimensional diffusion process with coefficient D and sampling it at random time intervals distributed exponentially with mean $\tau = 1/(Dm^2)$ (see Figure 3.2)

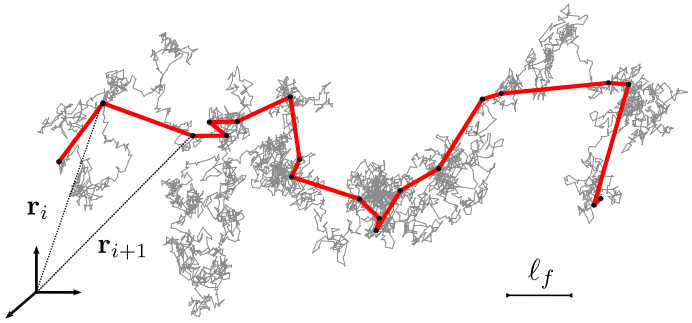


Figure 3.2: The red line is the free jump process where ℓ_f denotes the average jump length. The gray trajectory represents the Brownian motion related to the jump process. The red trajectory is obtained from the gray one sampling it with exponentially distributed time steps (black dots).

Consider a Brownian walker starting from $x_0 \in \mathbb{R}^3$ at time $t = 0$ and diffusing in a three-dimensional space with a diffusion coefficient D; its position x after a time t is given by the probability density function

$$G(x \,;\, t \,|\, x_0) = \frac{1}{(4\pi D t)^{3/2}} \, e^{-\frac{(x-x_0)^2}{4Dt}} \,. \tag{18}$$

We would like to introduce a characteristic length of the jump process. For this purpose, we sample the Brownian motion at exponentially distributed times with mean τ, thus defining a discrete random walk (see Fig.3.2).By definition, this new process is still Markovian and its successive positions x_i are given by the transition probability:

$$p(x_{i+1} \,|\, x_i) = \int_0^\infty \frac{\mathrm{d}t}{\tau} \, G(x_{i+1} \,;\, t \,|\, x_i) \, e^{-\frac{t}{\tau}} \,. \tag{19}$$

where τ characterises the coarse-graining. Passing to Fourier space we obtain

$$p(x_{i+1} \,|\, x_i) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \, \frac{m^2}{k^2 + m^2} \, e^{-i\, k \cdot (x_{i+1} - x_i)} \,, \tag{20}$$

where $m = 1/\sqrt{\tau D}$, which finally leads to Eq. (16). The jumps $(x_{i+1} - x_i)$ are identically distributed with mean length $\ell_f \doteq \langle ||x_{i+1} - x_i|| \rangle = 2/m$.

Let us now constrain this process to stay inside a cylindrical channel (see Fig. 3.3) of radius R and axial length $2H$, with the terminal domain $\mathcal{T}$ to be at one end of the cylinder

$$\mathcal{C} = \{x, y, z | \rho = \sqrt{x^2 + y^2} \leq R, |z| < H\}$$
$$\mathcal{T} = \{x, y, z | \rho = \sqrt{x^2 + y^2} \leq R, z > H\}.$$

Figure 3.3: $\mathcal{C}$ is the region where the trajectory has to be constrained. $\mathcal{T}$ is the target.

The choice of the Markov process described by (16) apart from being useful as an abstraction for a polymer it also helps in solving equation (8) when the domain $\mathcal{C}$ is an infinite cylinder and the terminal domain has been pushed to $z \to \infty$. The advantage comes from the fact that $p(x'|x)$ satisfies

$$(\boldsymbol{\nabla}^2 - m^2)\, p(x' \,|\, x) = -m^2\, \delta(x' - x) \,, \tag{21}$$

where δ is a Dirac delta function. Thus applying $\boldsymbol{\nabla}^2 - m^2$ to Eq. (8) gives

$$(\boldsymbol{\nabla}^2 - m^2)\, Z(x) = -m^2 \int_{\mathcal{C}} \mathrm{d}x' Z(x')\, \delta(x' - x) \,,$$

that leads to the system of equations

$$\begin{cases} \boldsymbol{\nabla}^2\, Z(x) = 0 \,, & \text{for } x \text{ in } \mathcal{C}, \\ (\boldsymbol{\nabla}^2 - m^2)\; Z(x) = 0 \,, & \text{elsewhere.} \end{cases} \tag{22}$$

Given that in cylindrical coordinates the problem is separable in z and (θ, ρ), the general solution of the equation above reads

$$Z(x) = \begin{cases} A\, \exp(\lambda z)\, J_0\left(\lambda \rho\right), & \text{for } x \text{ in } \mathcal{C} \\ B\, \exp(\lambda z)\, K_0\left(c_\lambda \rho\right), & \text{elsewhere,} \end{cases} \tag{23}$$

where A and B are real constants, $c_\lambda = \sqrt{m^2 - \lambda^2}$ and the parameter λ belongs to $(0, m)$ and satisfies

$$\begin{aligned} &\lambda\, J_1(\lambda R)\, K_0\left(\sqrt{m^2-\lambda^2} R\right) \\ &\qquad = \sqrt{m^2-\lambda^2}\, J_0(\lambda R)\, K_1\left(\sqrt{m^2-\lambda^2} R\right). \end{aligned} \tag{24}$$

In this expression, J_ν is Bessel functions of the first kind of oreder ν and K_ν is the modified Bessel function of the second kind of order ν.

This result leads to the reweighted transition probability in cylindrical coordinates $x = (\rho, \theta, z)$

$$q(x'|x) = \frac{J_0(\lambda \rho')}{J_0(\lambda \rho)}\, e^{\lambda(z'-z)} p(x'|x)\ . \tag{25}$$

for $\rho' < R$ and zero otherwise.

According to (25), λ^{-1} can be seen as a confinement-dependent length controlling the size distribution of the jumps in the positive $z-$direction: larger values of λ^{-1} will reflect in larger longitudinal jumps on average.
Figure 3.4 displays three realizations of the walk, when the reference jump length of the unconstrained walk, ℓ_f, increases with respect to the cylinder radius R, i.e. the system transitions from weak to strong confinement.

The knowledge of the transition probability (25) allows for a direct sampling of the constrained walk (see Appendix A.2 for details).

With this method, the complexity of generating trajectories is independent of the strength of the confinement, and grows linearly with the spanned size of the channel, and hence the average chain length. This allows us to produce large samples of confined trajectories without rejections. This is especially useful in the strong confinement limit (see Fig. 3.4), when the channel diameter is much smaller than the jumps of the unconstrained process. In this case, virtually all free trajectories will violate the constraints (see Appendix A.1). In analogy with the confined polymer in the following we discuss, several geometrical properties of interest. The result we obtained allows an efficient inspection of these properties going from very weak to strong confinement.

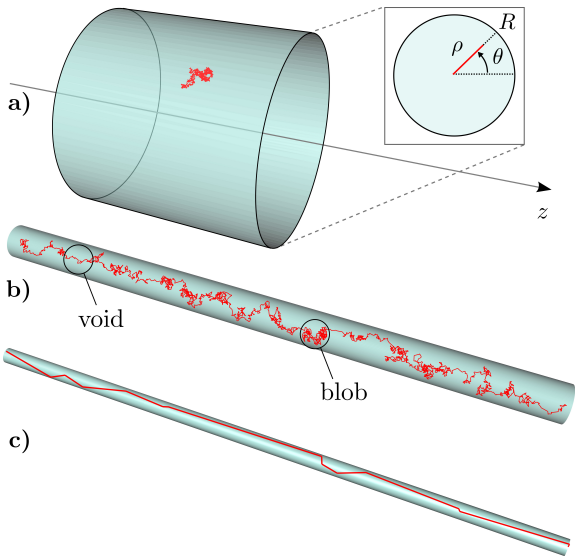


Figure 3.4: **Different regimes of confinement.** Monte Carlo simulations of a three-dimensional random walk confined in a cylindrical channel of axis z and radius R. *a) Weak confinement*, $\ell_f \approx 10^{-3}R$. The confinement has no effect on the jump process. *b) Intermediate confinement*, $\ell_f \approx 10^{-1}R$. The confinement starts to have an impact on the walk. Regions with higher and lower than average density of bonds appear, namely voids and blobs. *c) strong confinement*, $\ell_f \approx 2R$. In this regime the walker is force to make very straight jumps.

5 Geometrical properties of the confined chains

Exploiting the Monte Carlo algorithm explained in Appendix A.2 we studied several geometrical properties of the confined walks in analogy to what is usually done in polymer physics. But before showing the results we would like to spend few words to describe how these walks relates to polymers. The jump process in the constrained case corresponds to realizations of the free jump process that are entirely confined inside the cylinder, with one of their termini being anywhere inside the allowed cylindrical portion, $\mathcal{C}$, corresponding to $-H \leq z \leq H$ and $\rho < R$, while the other is in the absorbing domain $\mathcal{T}$, i.e. at $z > H$ and $\rho < R$. After performing the limit $H/R \to \infty$, this ensemble of trajectories can be thought as consisting of very long chains whose two termini (at $i \to \pm\infty$) are tethered at the opposite ends of the cylinder. We then look at the statistics of a subportion comprised between two tagged beads ($i = 0$ and $i = N$, see Fig. 3.5). Notice that this is a different ensemble from the one usually considered where the termini are unconstrained or for typical models of tensioned chains. For the latter, in fact, both ends are at the boundaries of the spanned region at sufficiently high tension, while in our ensemble only the last step has to do so. Another aspect we want to stress is that the average jump length varies as the degree of channel confinement increases. In standard polymer

models the bond length is, instead, kept fixed as the degree of spatial confinement or stretching is varied.

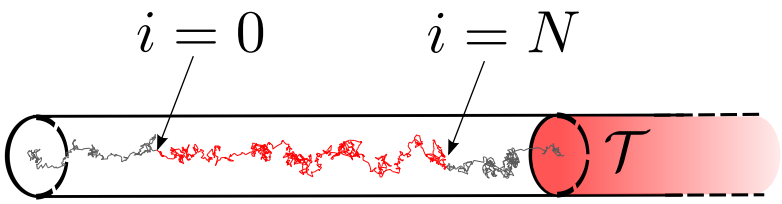


Figure 3.5: Eq. (20) is solved in the limit of an infinite cylinder. The ensemble we consider here can be thought as consisting of very long chains whose two termini (at $i \to \pm\infty$) are tethered at the opposite ends of the cylinder. We look at the statistics of the trajectory comprised between two tagged beads ($i = 0$ and $i = N$).

With these differences borne in mind, we investigated the behavior of several observables that have a straightforward interpretation in the language of polymer physics: the mean extension of the polymer along the z-axis, $L_z = \langle z_N - z_0 \rangle$, the end-to-end distance $R_{ee} = \sqrt{\langle \|x_N - x_0\|^2 \rangle}$ and the chain length $L = \langle \sum_{i=0}^{N-1} \|x_{i+1} - x_i\| \rangle$. Note that, as we discuss later, the conformational selection due to confinement is expected to make the average bond lengtho f the confined jump process, ℓ_c, larger than the free case one, ℓ_f.

We have investigate the previous properties for different values of ℓ_f, ranging from $\ell_f = 9 \cdot 10^{-5} R$ to $\ell_f = 3R$.

Three distinct regimes are observed (see Fig. 3.4):

a) *weak confinement*, where the average bond length of the constrained walk is still comparable to the free one, ℓ_f, and $R_{ee} \ll R$. In this regime the metric properties of the chain are only slightly perturbed with respect to the free case;

b) *intermediate confinement*, where the average bond length is still small relative to the cylinder diameter but the end-to-end distance is comparable or larger than it ($\ell_c \sim \ell_f \ll R \lesssim R_{ee}$);

c) *strong confinement*, where the chain is affected even at the scale of individual bonds, $\ell_c \gg \ell_f \simeq R$.

In the left panel of Fig. 3.6 we show the scaling behavior of the end-to-end distance as a function of the chain length. The blue curves are in the weak confinement case while the red ones are for the strong confinement, while the yellow ones (and the black circles) correspond to the intermediate confinement regime. The emergence of two different scaling regimes becomes evident upon rescaling the chain length, $L \to \lambda L/m$, which produces a nearly exact collapse of the data (see the left panel of Fig. 3.6). Note that in the strong confinement regime, the end-to-end distance

is proportional to the total chain length. This is reminescent of the Odijk scaling regime for polymers that are strongly confined inside channels[18–20].

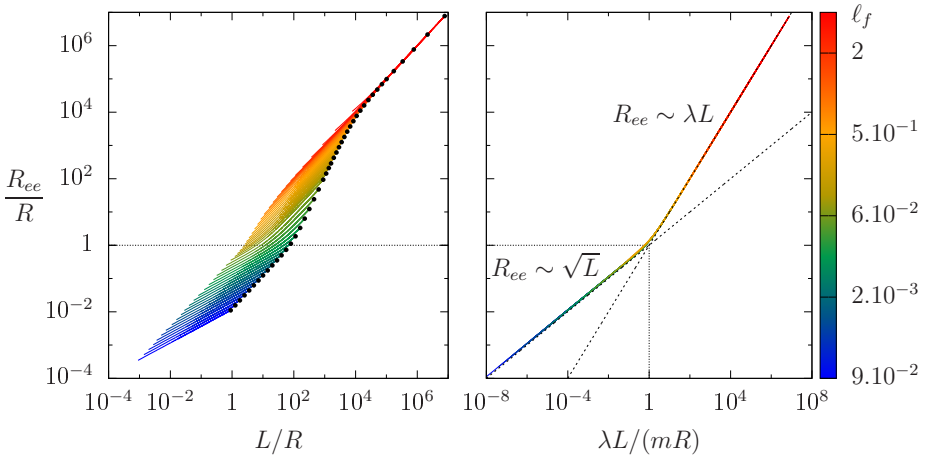


Figure 3.6: **$\mathbf{R_{ee}}$ behavior from weak to strong confinement.** *Left panel.* R_{ee} as a function of L. Each colored curve is obtained as an average of 10^4 realizations of the walk with a fixed value of the parameter ℓ_f. *Right panel.* Same as for the left panel, after rescaling $L \to \lambda L/m$.

In the left panel of Fig. 3.7 we show the deviation of the average length ℓ_c of the segments of the confined walk from the average length ℓ_f in the free case as a function of the strength of the confinement $2R/\ell_f$. Under weak confinement ($\ell_f \ll R$), ℓ_c is essentially equal to the mean length ℓ_f of the bonds for the free polymer, whereas under strong confinement ($\ell_f \gtrsim R$), ℓ_c increases exponentially fast (see Appendix A.6). This behavior signals the appearance of long stretches of nearly linear polymer configurations in the limit of strong confinement when the free average bond length ℓ_f exceeds the channel radius R.

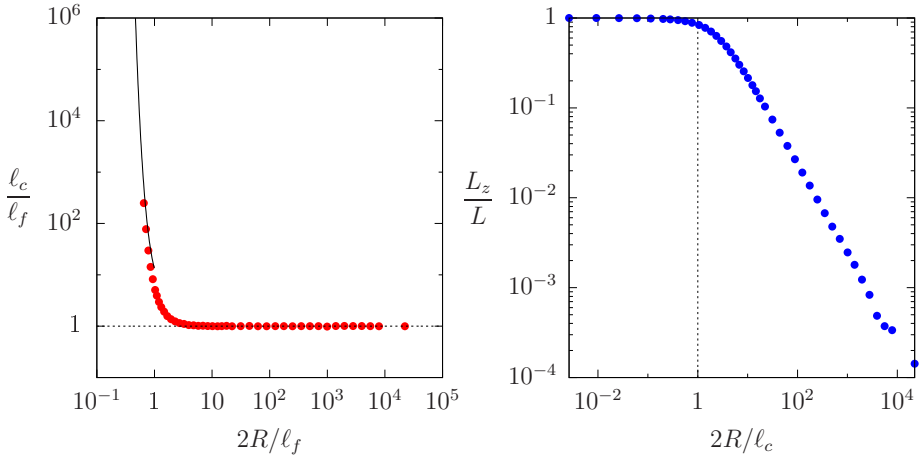


Figure 3.7: **Effect of the confinement on the average jump.** *Left panel*: The ratio beetwen the average jump length ℓ_c of the confined walker and the free jump length as a function of the parameter $2R/\ell_f$ (red). For $\ell_f \ll 2R$, ℓ_c is essentially equal to the free jump size ℓ_f (dashed line), whereas for $\ell_f \gtrsim 2R$ it increases exponentially fast. The black line is its analytical expression in this regime (see Appendix A.6). *Right panel*: The extension L_z/L as a function of $2R/\ell_c$. Each value of L_z and L is obtained over 10^4 realizations of the walk with fixed parameter ℓ_f.

Another customary way to present results on the elongation statistics is given in the right panel of Figure 3.7. It displays L_z/L for different values of $2R/\ell_c$ expressing the degree of confinement. We observe two main regimes, for $2R/\ell_c$ smaller or larger than 1. In the diffusive limit ($\ell_f \sim \ell_c \ll R$), L_z/L varies linearly with ℓ_c/R, with a slope that can be computed analytically in the diffusive limit (see Appendix A.5).

6 Fluctuations in the density of bonds along the channel

Visual inspection of the confined paths (see Fig. 3.4) suggests that they are not homogeneously dense along the channel but rather feature an alternation of densely and sparsely occupied regions. To quantify this effect, we considered a measure of the variations of the local density of bonds along the channel. This is defined as the number of bonds that fall inside a cylindrical region of width Δ along the z-axis. In particular we ask how this density deviates from a Poisson distribution, which is the reference for a uniformly distributed point process. We selected a set of parameters corresponding to an intermediate confining situation (see Fig. 3.4b). By projecting the trajectory along the z-axis we build the empirical probability distribution to have n points inside a region of width Δ. To explicitly highlight the presence of regions denser than average we define the rescaled variable $(n - \langle n \rangle)/\sigma_n$, where $\langle n \rangle$ and σ_n are respectively the mean and the standard deviation of the number of points n for

a given width Δ.

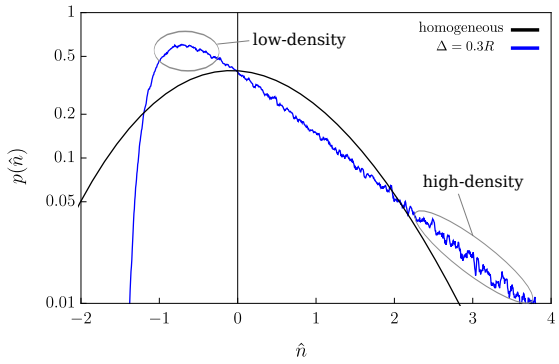


Figure 3.8: **Voids and blobs.** The curves represent the probability density $p(\hat{n})$ for $\hat{n} = (n - \langle n \rangle)/\sigma_n$ in the homogeneous case (black curve) and for $\Delta = 0.3R$ (blue curve) in the case $\ell_f/R \approx 10^{-2}$.The encircled regions highlight inhomogeneities in the density of bonds. The right tail indicates the presence of regions with higher density of points and the peak at negative values stems from the presence of regions with lower density of points. Plots for different values of Δ and comparison with analytical results are shown in Appendix A.9.

Fig. 3.8 clearly highlights the presence of local inhomogeneities in the distribution. In fact, the shape of the right tail suggests that in the confined process there is a higher probability to have regions with higher than average number of points. The location of the peak of the distribution also shows that less dense regions are more likely in the confined case. As detailed in Appendix A.9 the shape of these distributions is well captured by the diffusive approximation for the confined process.

7 Conclusions and perspectives

In this Chapter we proposed a general framework for the exact and efficient generation of constrained random walks. The formalism in its full generality can be applied to all Markovian jump processes. In general, one has to solve numerically the linear equation (8) and use its solution to obtain the transition probability for the constrained process which can then be directly sampled by any suitable technique. Sometimes it is possible to obtain exactly the transition probability and gain significant analytical control on the process and exceptional efficiency.

Inspired by the classical problem of polymer chains confined inside nano-channels, which are still actively investigated for their rich metric and entanglement properties, we have applied this method to a jump process constrained inside a cylinder.

For the purposedly minimalistic jump processes considered here, we have shown that the proposed strategy offers an effective way of implementing confining constraints that would otherwise make the problem intractable with simple sampling

(rejection-based) sampling strategies.

We also highlighted that the confined jump process is exactly equivalent to a Markov decision process. In this respect we were able to solve the optimal Bellman equation and find the optimal policy achieving the goal of reaching a given target respecting a geometrical constraint. One of the properties of polymer models is self-avoidance (i.e. the bonds cannot cross each other). The cost function we analyzed here do not take into account any interaction cost between the jumps (i.e. the walks in this case is not self-avoiding) and it would be very interesting to investigate cases in which this interaction is taken into account allowing, if possible, to generate configurations closer to self-avoiding polymers.

Following the analogy with search strategies, the agent perfectly knows its position in the space of states (here the physical space) and has perfect representation of where the target is located.

In the next Chapter we study an example inspired by bird navigation where the knowledge of the environment and the target are both represented by partial knowledge.

Appendix A

A.1 Generating the constrained ensemble by rejection

Directly sampling a free random walk with exponentially distributed jumps according to eq. (16) and then rejecting the trajectories that do not satisfy the constraint of staying within a cylinder is very inefficient, especially for long chains and large jumps. Figure 3.9 shows the efficiency of this brute-force method for different number of jumps and their length. The efficiency falls off exponentially fast to zero as expected. Fig. 3.10 presents a summary of the results concerning the runtime to sequentially produce one hundred trajectories using the rejection method and the method described in this work. In the case of small jump length the rejection method is impractical if one wants to explore the limit of long chains (see Fig. 3.10 right panel). The other limit is shown in the left panel where the jump length is comparable with the confinement. In this case sampling few hundreds of jumps is impractical. Using the rejection method, the runtime grows exponetially with N in agreement with what shown in Fig. 3.9. The method we described in this work instead presents a runtime that practically grows linearly with N.

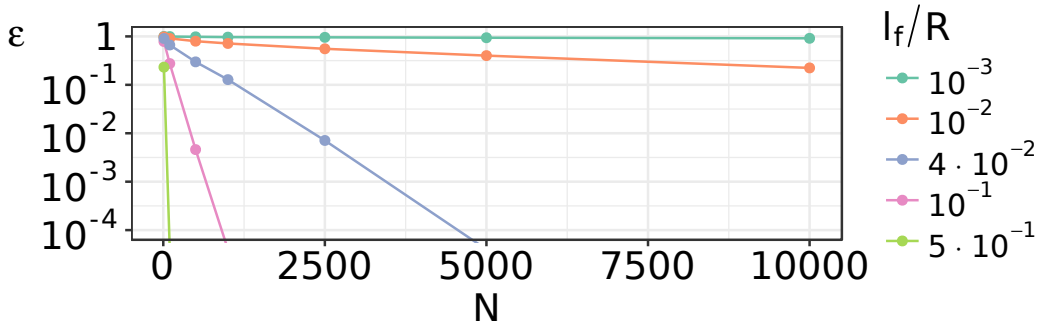


Figure 3.9: Fraction of free chains that satisfy the constraint as a function of the number of jumps. Colors encode the jump length and N is the number of jumps.

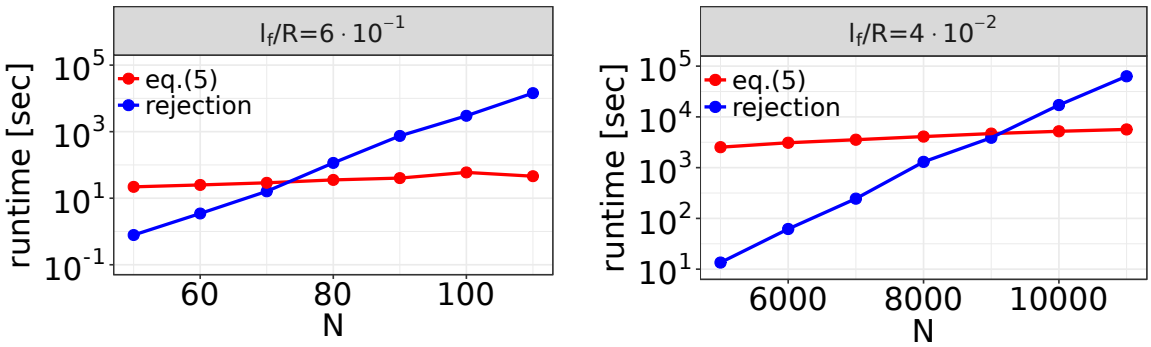


Figure 3.10: The CPU time required to generate ensembles of 100 paths of length N using the strategy based on Eq. (25) (red points) is compared with the one required by the rejection method (blue points). The two panels correspond to different degrees of channel confinement: $\ell_f/R = 6 \cdot 10^{-1}$ and $\ell_f/R = 6 \cdot 10^{-1}$ for the left and right panels, respectively. In both cases, the runtime grows linearly with N for the the strategy based on Eq. (25), while it grows exponentially with N for the rejection method. As the chain length is increased, this significant additional computational cost make the rejection scheme impractical compared to the proposed strategy.

A.2 Monte Carlo simulation of confined trajectories

As the transition probability (25) is invariant under translation along the axis z, the distribution of the walker positions in the transverse direction reaches a stationary state:

$$P_{st}(\rho,\theta) = \frac{J_0^2(\lambda\rho)}{\pi\, R^2\,[J_0^2(\lambda R) + J_1^2(\lambda R)]}\,, \tag{A.1}$$

that verifies, for all $\rho < R$,

$$\int_0^R \rho' \mathrm{d}\rho' \int_0^{2\pi} \mathrm{d}\theta'\, q(\rho,\theta\,|\,\rho',\theta')\, P_{st}(\rho',\theta') = P_{st}(\rho,\theta)\,,$$

where

$$\int_0^{2\pi} \mathrm{d}\theta' q(\rho,\theta\,|\,\rho',\theta') = \int_{-\infty}^{+\infty} \mathrm{d}z \int_0^{2\pi} \mathrm{d}\theta'\, q(\rho,\theta,z\,|\,\rho',\theta',0)\,,$$

$$= m^2\, \frac{J_0(\lambda\rho)}{J_0(\lambda\rho')} \begin{cases} I_0(c_\lambda\rho')\, K_0(c_\lambda\rho) & \text{if } \rho > \rho' \\ I_0(c_\lambda\rho)\, K_0(c_\lambda\rho') & \text{else.} \end{cases}$$

Starting from steady state (A.1) in the transverse direction of the channel, we then generate confined trajectories using a direct Monte Carlo method with the jump

process (25). To sample jumps from (25), we use the acceptance-rejection method with the instrumental pdf

$$f(x_1|x_0) = \frac{m^2\, e^{-m\,||x_1-x_0||}}{4\pi\,||x_1-x_0||\,C(x_0)}\, e^{\lambda(z_1-z_0)}\, \mathbb{I}_{\rho_1<R}\,, \tag{A.2}$$

where $C(x_0) = m^2\,[1 - c_\lambda R\, I_0(c_\lambda\rho_0)\, K_1(c_\lambda R)]\,/c_\lambda^2$. The cumulative distribution function (cdf) of Eq. (A.1) can be exactly computed thus sampling ρ does not represent a computational bottleneck. The cdf for (A.2) cannot be expressed in a closed form but this pdf is indeed simpler to sample due to the absence of the Bessel functions of the first kind. For all x_0 and x_1 in the cylinder, $q(x_1\,|\,x_0) \le k\, f(x_1\,|\,x_0)$, where $k = C(\rho_0)/J_0(\lambda\rho_0)$, which sets the rejection threshold to

$$0 < \frac{q(x_1\,|\,x_0)}{k\, f(x_1\,|\,x_0)} = J_0(\lambda\rho_1) \le 1. \tag{A.3}$$

Observe that this threshold decreases when the walk gets closer to the boundaries ($\rho_1 \to R$) and closer to the diffusion limit ($\lambda R \to z_{0,1}$). It finally remains to show how to sample from $f(x_1\,|\,x_0)$. Using the change of variables

$$\begin{cases} \ell = ||x_1 - x_0||\,, \\ \xi = \cos\nu = \dfrac{z_1 - z_0}{\ell} \;\in [-1,1]\,, \\ \tan\varphi = \dfrac{y_1 - y_0}{x_1 - x_0}\,, \quad \varphi \in [-\pi,\pi]\,, \end{cases} \tag{A.4}$$

we can rewrite the pdf (A.2) as

$$f(x_1|x_0)\,\mathrm{d}^3x_1 = \mathbb{I}_{\ell<\ell^*}\, \frac{m^2\, e^{-m\,\ell+\lambda\,\ell\,\xi}}{4\pi\, C(\rho_0)}\, \ell\,\mathrm{d}\ell\,\mathrm{d}\xi\,\mathrm{d}\varphi\,, \tag{A.5}$$

where $\ell^*(x_0,\xi,\varphi)$ is the maximum length that a walker starting from x_0 can travel within the cylinder in the direction given by (ξ,φ): $\ell^* = b(x_0,\varphi)/\sqrt{1-\xi^2}$, with $b(x_0,\varphi) = \sqrt{R^2 - \rho_0^2\sin^2(\varphi-\theta_0)} - \rho_0\cos(\varphi-\theta_0)$. The joint probability density can then be decomposed into three probability densities, one for each variable:

$$f(\ell,\,\xi,\,\varphi\,|x_0)\,\ell^2 = f(\ell\,|\,\xi,\,\varphi,\,x_0) f(\xi\,|\,\varphi,\,x_0) f(\varphi\,|\,x_0)$$

where

$$\begin{cases} f(\varphi\,|\,x_0) &= \int_0^{+\infty}\ell^2\mathrm{d}\ell\int_{-1}^{1}\mathrm{d}\xi\, f(\ell,\,\xi,\,\varphi\,|x_0) \\ f(\xi\,|\,\varphi,\,x_0) &= \dfrac{\int_0^{+\infty}\ell^2\mathrm{d}\ell\; f(\ell,\,\xi,\,\varphi\,|x_0)}{f(\varphi\,|x_0)} \\ f(\ell\,|\,\xi,\,\varphi,\,x_0) &= \dfrac{\ell^2\, f(\ell,\xi,\varphi\,|x_0)}{f(\xi\,|\,\varphi,x_0)\, f(\varphi\,|x_0)}\,. \end{cases}$$

Starting from x_0, we thus sample first the angle φ from

$$f(\varphi) = \frac{m^2}{2\pi\, c_\lambda^2\, C(\rho_0)} \left\{1 - c_\lambda\, b(\varphi)\, K_1[c_\lambda b(\varphi)]\right\}\,, \tag{A.6}$$

then, given the angle φ, we then obtain ξ from the pdf

$$f(\xi\,|\,\varphi) = \frac{m^2}{4\pi\, C(\rho_0)\, f(\varphi)} \frac{1 - e^{-(m-\lambda\xi)\ell^*}[1 + (m-\lambda\xi)\ell^*]}{(m-\lambda\xi)^2}.$$

Finally, the jump length ℓ is sampled directly from

$$f(\ell\,|\,\xi,\varphi) = \frac{(m-\lambda\,\xi)^2 \quad \ell\, e^{-(m-\lambda\xi)\ell}\, \mathbb{I}_{\ell\le\ell^*}}{1 - e^{-(m-\lambda\xi)\ell^*}\,[1 + (m-\lambda\xi)\ell^*]}\,, \tag{A.7}$$

using the Lambert W function.

A.3 Constitutive equation for λ

In this section we derive equation (24) that defines λ. This is a parameter biasing the average jumps along the $z-$axis and must be determined just once at the beginning of each simulation.

In cylindrical coordinates $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ is expressed by (ρ, ϕ, z) so that $Z(\mathbf{x}) = A\exp(\lambda z)J_0(\lambda\rho)$ with $\rho = \sqrt{x^2+y^2}$. Using (20) and integrating over z equation (8) takes the form

$$\exp(\lambda z)J_0(\lambda\rho) = m^2 e^{\lambda x_1} \int \frac{d\kappa}{2\pi}\frac{d\varphi}{2\pi}\kappa \frac{e^{-i\kappa\rho\cos\varphi}}{\kappa^2+m^2-\lambda^2} \int d\rho' d\varphi' \rho' e^{i\kappa\rho'\cos\varphi'} J_0(\lambda\rho')$$

The right hand side after integration over ϕ and ϕ' is

$$\texttt{r.h.s} = m^2 e^{\lambda z} \int_0^\infty d\kappa \frac{\kappa J_0(\kappa\rho)}{\kappa^2+m^2-\lambda^2} \int_0^R d\rho'\rho' J_0(\kappa\rho')J_0(\lambda\rho') \tag{A.8}$$

Rewriting $\frac{m^2}{\kappa^2+m^2-\lambda^2} = 1 - \frac{\kappa^2-\lambda^2}{\kappa^2+m^2-\lambda^2}$ the previous integral becomes

$$\begin{aligned}\texttt{r.h.s} = e^{\lambda z}\Big[&\int_0^R d\rho'\rho' J_0(\lambda\rho') \int_0^\infty d\kappa\kappa J_0(\kappa\rho)J_0(\kappa\rho') \quad \text{Ⓐ} \\ &- \int_0^\infty d\kappa\kappa J_0(\kappa\rho) \int_0^R d\rho'\rho' J_0(\lambda\rho' J_0(\kappa\rho') \frac{\kappa^2-\lambda^2}{\kappa^2+m^2-\lambda^2}\Big] \quad \text{Ⓑ}\end{aligned}$$

Thanks to the closure formula for the Bessel functions

$$\int_0^\infty d\kappa\kappa J_0(\kappa\rho)J_0(\kappa\rho') = \delta(\rho-\rho')/\rho'$$

$$\text{Ⓐ} = e^{\lambda z} J_0(\lambda\rho)$$

Then in Ⓑ we can perform the integral in ρ' (Lommel's Integrals) that results in

$$\text{Ⓑ} = \int_0^\infty d\kappa\kappa J_0(\kappa\rho)\frac{\kappa J_0(\lambda R)J_1(\kappa R) - \lambda J_0(\kappa R)J_1(\lambda R)}{\kappa^2 + m^2 - \lambda^2} = 0. \tag{A.9}$$

The integral over κ is known (Gradshteyn Ryzhik 6.577 2) leading to

$$\frac{J_1(\lambda R)}{J_0(\lambda R)} = \frac{\sqrt{m^2-\lambda^2}K_1(R\sqrt{m^2-\lambda^2})}{\lambda K_0(R\sqrt{m^2-\lambda^2})}. \tag{A.10}$$

that is equation (24).

A.4 Asymptotic Behavior of λR

The left panel of Fig. 3.11 displays the behavior of λR as a function of ℓ_f/R obtained by solving numerically Eq. (24). In the (diffusive) limit $\ell_f/R \ll 1$, λ behaves as

$$\lambda\, R \sim z_{0,1}\left(1 - \frac{\ell_f}{2\,R}\right), \tag{A.11}$$

where $z_{0,1} \simeq 2.40483$ is the first zero of the Bessel function J_0. In the limit $\ell_f/R \gg 1$, Eq. (24) gives the asymptotic behavior

$$\lambda R \sim \frac{2R}{\ell_f}\left[1 - \frac{1}{2}\left(\frac{\ell_f}{2R}\right)^2 \exp\left[-4\left(\frac{\ell_f}{2R}\right)^2\right]\right], \tag{A.12}$$

using that, in this limit, $\max(\lambda R, \sqrt{m^2-\lambda^2}R) \leq mR \ll 1$ and $c_\lambda R \sim \exp[-2/(mR)^2]$. Note that $\lambda\, R \in (0, z_{0,1})$ and is a strictly decreasing function of ℓ_f/R.

A.5 Theoretical analysis of the polymer extension $L_z = \langle z_N - z_0 \rangle$

After relaxation to steady state in the transverse direction (see Appendix 5), the generating function of the jump lengths $z = (z_{i+1} - z_i)$ along the axis of the cylinder is given by

$$G(s) = \langle e^{sz} \rangle = \int_{-\infty}^{+\infty} \mathrm{d}z\; e^{sz}\; p(z)\,, \qquad \text{where} \tag{A.13}$$

$$p(z) = \int \rho\,\mathrm{d}\rho\,\mathrm{d}\theta \int \rho'\mathrm{d}\rho'\,\mathrm{d}\theta'\; q(\rho', \theta', z\,|\,\rho, \theta, 0)\, P_{st}(\rho, \theta)\,.$$

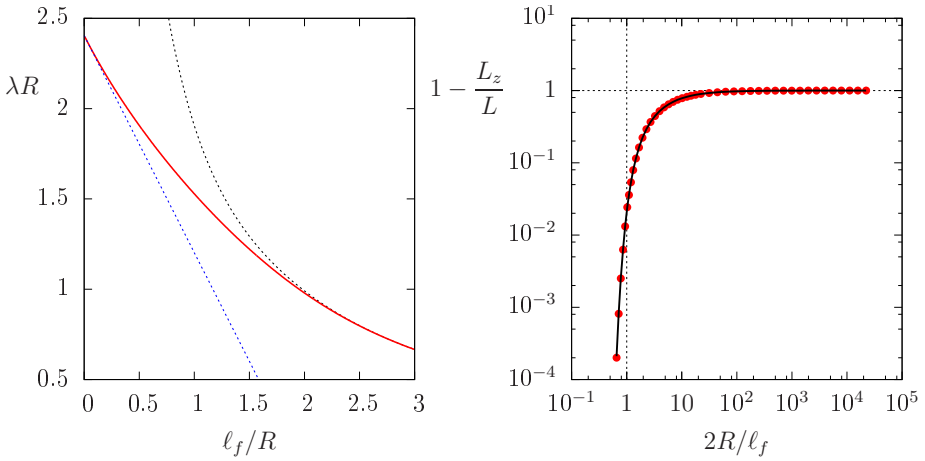


Figure 3.11: *Left panel:* Values of λR (red) obtained by inverting numerically Eq. (24) for different values of ℓ_f/R. The two asymptotic curves are also displayed: in blue, Eq. (A.11) for $\ell_f/R \ll 1$, and in black, Eq. (A.12) for the long jump limit $\ell_f/R \gg 1$. *Right panel:* Numerical estimates of $1 - L_z/L$ for different values of ℓ_f/R (red dot), compared with the curve $1 - \lambda\ell_f/2$ (in black).

Integrating, we obtain

$$G(s) = \frac{m^2}{m^2 - s^2 - 2\lambda s} \times \left[1 + 2\,\frac{[\lambda\, J_1K_0 - c_\lambda(s)\, J_0K_1]\;\; [\lambda\, J_1I_0 + c_\lambda(s)\, J_0I_1]}{(J_0^2 + J_1^2)\;(m^2 - s^2 - 2\lambda s)}\right], \tag{A.14}$$

where $c_\lambda(s) = \sqrt{m^2 - (\lambda + s)^2}$, $K_i = K_i[c_\lambda(s)\, R]$, $I_i = I_i[c_\lambda(s)\, R]$ and $J_i = J_i(\lambda R)$. Finally expanding to first order at $s = 0$ leads to the expression of the mean jump length along the z-axis, $\ell_z = \langle z\rangle = G'(s)|_{s=0}$:

$$\ell_z = \frac{\lambda\ell_f^2}{2}\left[1 + g\left(\frac{\ell_f}{2R}\right)\right], \tag{A.15}$$

where g is a positive and increasing function of ℓ_f/R that vanishes as ℓ_f/R goes to 0:

$$g\left(\frac{\ell_f}{2R}\right) = \frac{\lambda J_1I_0 + c_\lambda J_0I_1}{c_\lambda\,(J_0^2 + J_1^2)}\,[J_0(2K_1 - c_\lambda RK_2) + \lambda RJ_1K_1] \tag{A.16}$$

with $J_i = J_i(\lambda R)$, $K_i = K_i(c_\lambda R)$ and $I_i = I_i(c_\lambda R)$. We recall that both λR and $c_\lambda R$ are functions of $mR = 2R/\ell_f$. In the diffusive limit $\ell_f \ll 2R$, $\lambda\ell_f$ behaves as $O(\ell_f/R)$ (see Eq. (A.11)), and expanding (A.15) to first order thus yields the asymptotic behavior:

$$\ell_z = \ell_f\left[\frac{\lambda\ell_f}{2} + o\left(\frac{\ell_f}{2R}\right)\right] \sim z_{0,1}\,\frac{\ell_f^2}{2R}\,. \tag{A.17}$$

This result is in agreement with the direct derivation for the diffusive case, where we found the drift $2D\lambda$ along the cylinder axis (see Appendix A.7). Since the process is Markovian and is started from steady state, the mean extension of the polymer can be decomposed as $L_z = N\ell_z$, and the total length of the polymer as $L = N\ell_c$ (see main text). Moreover, in the diffusive limit, $\ell_c \sim \ell_f$, so that Eq. (A.17) becomes

$$\frac{L_z}{L} = \frac{\ell_z}{\ell_c} \sim \lambda \frac{\ell_f}{2} \qquad \text{with} \;\; \lambda \sim \frac{z_{0,1}}{R}\,, \tag{A.18}$$

which varies linearly with ℓ_f/R (see left panel of Fig. **??**). Note that, in the long-jump limit $mR \ll 1$, the asymptotic expansion of Eq. (A.16) is

$$g(mR) = \frac{(mR)^4}{4} \exp\left[\frac{4}{(mR)^2}\right] - 1 + O(mR)^2\,. \tag{A.19}$$

A.6 Numerical analysis

Polymer extension L_z – Perhaps surprisingly, we observe from the numerical simulations that Eq. (A.18) extends to any value of ℓ_f,

$$\text{for any } \ell_f, \qquad \frac{L_z}{L} = \frac{\ell_z}{\ell_c} \simeq \lambda \frac{\ell_f}{2}\,, \tag{A.20}$$

where λ is now given by Eq. (24), as shown in the Right panel of Fig. 3.11 where Eq. (A.20) matches almost perfectly the numerical data. In the long-jump regime $\ell_f \gg R$, using the expansion (A.12) for λR, we obtain

$$\text{for } \ell_f \gg R, \quad \frac{L_z}{L} \sim 1 - \frac{1}{2}\left[\left(\frac{\ell_f}{2R}\right) e^{-2\left(\frac{\ell_f}{2R}\right)^2}\right]^2. \tag{A.21}$$

Mean jump length under confinement ℓ_c – Reformulating Eq. (A.20) with Eq. (A.15) we can write:

$$\text{for any } \ell_f, \qquad \ell_c \simeq \ell_f \left[1 + g\left(\frac{\ell_f}{2R}\right)\right], \tag{A.22}$$

which stays consistent with the numerical data (see Fig. 3.7). In the diffusive limit we recover that $\ell_c \sim \ell_f$, and, in the long-jump limit, using Eq. (A.19), we can note the extremely rapid growth of ℓ_c as ℓ_f/R increases:

$$\text{for } \;\ell_f \gg R, \qquad \frac{\ell_c}{2R} \sim \frac{1}{4}\left(\frac{2R}{\ell_f}\right)^3 e^{4\left(\frac{\ell_f}{2R}\right)^2}. \tag{A.23}$$

End-to-end distance $R_{ee} = \sqrt{\langle \|x_n - x_0\|^2 \rangle}$ – In Fig. 3.6, each colored curve (fixed value of ℓ_f/R) displays two distinct regimes: $R_{ee} \propto \sqrt{L}$ for $R_{ee} \ll R$ and $R_{ee} \propto L$ for $R_{ee} \gg R$. For polymer such that $R_{ee} \ll R$, the polymer behavior can be modeled by a Brownian walker under confinement, which yields,

$$R_{ee}^2 = 6Dt + (2\lambda Dt)^2 \,, \tag{A.24}$$

with $t = N\tau = \ell_f L/4$ and where the second term comes from the drift along the z-axis resulting from the confinement (see Appendix A.7). Replacing the value of t, we observe that this second term is negligible with respect to the first one, so that R_{ee} evolves as $\sqrt{6Dt}$:

$$\text{for } R_{ee} \ll R, \qquad \frac{R_{ee}}{R} \sim \sqrt{3\,\frac{\ell_f}{2R}\,\frac{L}{R}}\,; \tag{A.25}$$

the effect of the confinement on R_{ee} is not visible in this regime. For long polymers $R_{ee} \gg R$, we observe that R_{ee} can be rescaled using $L \to \lambda\,\ell_f L/2$ (see Fig. 3.6). Note that this rescaling is valid only after a large number of jumps, $N \gg 1$, and is not exact. Indeed, by definition we can write $R_{ee} = \sqrt{\langle L_\perp^2 \rangle + \langle\,(z_N - z_0)^2\rangle}$ where $L_\perp$ in defined in the transverse direction of the cylinder. The rescaling does not apply to $\langle L_\perp^2 \rangle$ in the long-polymer regime (see Fig. 3.12). However this is not visible on R_{ee}, as, for long polymer ($R_{ee} \gg R$), $\langle\,(z_N - z_0)^2\rangle \gg \langle L_\perp^2 \rangle$, and therefore $R_{ee} \sim \sqrt{\langle\,(z_N - z_0)^2\rangle}$ that rescales for large N.

Now keeping the number of jumps N fixed and varying the effective channel size (see black dots in Fig. 3.6), we observe three main regimes. They result from the overlap of the two transitions previously described, for R_{ee} and L_z, and are summarized in Fig. 8.

A.7 Brownian motion constrained inside a cylinder

In the continuum ($l_f/R \to 0$) the jump process becomes a controlled Brownian motion for which an analytical description is affordable. The effect of confining a Wiener process in the cylindrical channel $\mathcal{C}$ is subsumed by an additional drift term [10], $u(x)$, called control drift. The Langevin equation for the walker thus reads

$$\dot{x}(t) = \mathbf{u}(x) + \sqrt{2D}\,\boldsymbol{\eta}_t \,, \tag{A.26}$$

where each η_t^i is an independent white noise. The behaviour of the confined Brownian process then corresponds to the optimal (stationary) trajectories of this controlled walker: we look for the optimal control drift using the Hamilton-Jacobi-Bellman

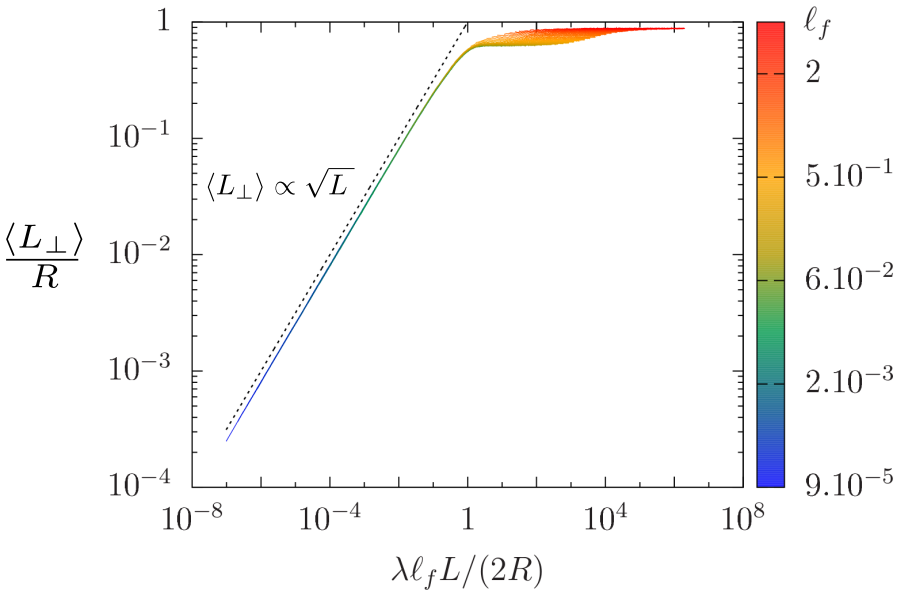


Figure 3.12: Evolution of $\langle L_\perp \rangle / R$ as a function of the rescaled variable $\lambda \ell_f L/(2R)$ in log-log scale. Each colored curve is obtained from 10^4 realizations of the walk with a fixed value of the parameter ℓ_f (see scale on the right), varying the total length of the polymer (from $N = 1$ to $N = 10^4$ jumps). Observe that the rescaled curves do not collapse for values of $\langle L_\perp \rangle / R$ close to 1.

equation with a cost that takes into account the boundaries (see [11]). We find that the drift $\mathbf{u}(x)$ takes the form:

$$\mathbf{u}(\rho) = 2D\lambda\, \mathbf{e}_z - 2D\lambda\, \frac{J_1(\lambda\,\rho)}{J_0(\lambda\,\rho)}\, \mathbf{e}_\rho \ , \tag{A.27}$$

where $\lambda = z_{0,1}/R$. As a consequence, the mean length travelled by the confined Brownian walker in the direction of the z-axis during a time τ is

$$\ell_z = \langle z \rangle = 2D\lambda\,\tau \, . \tag{A.28}$$

For the process (19), where $\tau D = \ell_f^2/4$, we thus expected to recover, in the diffusive limit, that $\ell_z = \lambda\,\ell_f^2/2$. For the same reason, we find that the mean-square distance travelled in the direction of the z-axis during a time t is given by:

$$\langle\, z^2(t)\,\rangle = \sigma(t)^2 + \langle z(t)\rangle^2 = 2D\,t + (2D\,\lambda\,t)^2 \, . \tag{A.29}$$

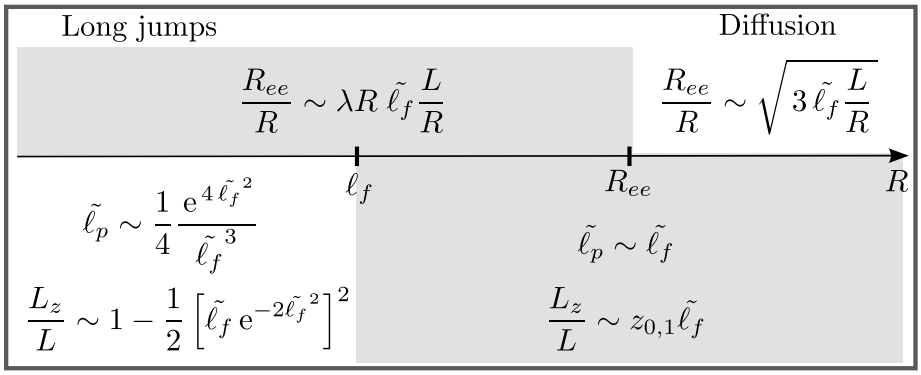


Figure 3.13: Summary of the results found analytically and numerically, for L_z and ℓ_c in the regimes $R \ll \ell_f$ and $R \gg \ell_f$, and for R_{ee} in the regimes $R \ll R_{ee}$ and $R \gg R_{ee}$. We took the notations $\tilde{\ell_f} = \ell_f/(2R)$ and $\tilde{\ell_c} = \ell_c/(2R)$. General expressions for L_z/L and $\tilde{\ell_c}$, found valid in any regime, are given, respectively, in Eq. (A.20) and Eq. (A.22).

A.8 The limit $H/R \to \infty$ for the constrained Brownian motion

The Laplace equation in cylindrical coordinates is

$$\boldsymbol{\nabla}^2 Z(\rho,\, \theta,\, z) = \frac{1}{\rho}\partial_\rho\left(\rho\, \partial\rho Z\right) + \frac{1}{\rho^2}\partial_\theta^2 Z + \partial_z^2 Z = 0$$

The cylinder has a radius R and in the z direction it extends from $-H$ to H. We impose the following boundary conditions for the Laplace equation:

$$Z(R,\, \theta,\, z) = 0 = Z(\rho,\, \theta,\, -H)$$
$$Z(\rho,\, \theta,\, H) = 1$$

The equation is separable and, looking for a solution of the kind $Z(\rho,\, \theta,\, z) = P(\rho)\, \Theta(\theta)\, \zeta(z)$, it can be written as the following equivalent system of coupled ordinary differential equation:

$$\zeta''(z) = \lambda^2\, \zeta(z),$$
$$\Theta''(\theta) = -\mu^2\, \Theta(\theta),$$
$$\rho^2 P''(\rho) + \rho P'(\rho) + (\lambda^2\rho^2 - \mu^2)P(\rho) = 0,$$

where here λ and μ are real parameters. The solution to the equation for ζ which satisfies the Dirichlet boundary conditions on the left end of the cylinder is

$$\zeta(z) = const \times \sinh[\lambda(z + H)]$$

The equation for Θ satisfying the rotational invariance about the longitudinal axis of the cylinder selects the value $\mu = 0$ and is just a constant:

$$\Theta(\theta) = const$$

Finally, the solution for P is the regular Bessel function of first kind of order zero:

$$P(\rho) = const \times J_0(\lambda\,\rho)$$

the allowed values of λ are all and only those for which $P(R) = 0$, so $\lambda_n = z_{0,n}/R$, where we denote by $z_{0,n}$ the n-th zero of $J_0(x)$.
Therefore, the solution of the Laplace equation in the cylindrical geometry specified above is, dropping the θ dependence,

$$Z(\rho,\, z) = \sum_{n=1}^{\infty} c_n\, \sinh[z_{0,n}\,(z+H)/R]\, J_0(z_{0,n}\rho/R)$$

The vanishing conditions at $\rho = R$ and $z = -H$ is already implemented in the solution, while the boundary condition $Z|_{z=H} = 1$ fixes the coefficients c_n as the solution of

$$\sum_{n=1}^{\infty} \tilde{c}_n\, J_0(z_{0,n}x) \equiv \sum_{n=1}^{\infty} \tilde{c}_n\, J_{0,n}(x) = 1 \qquad \forall\, x = \frac{\rho}{R} \in [0,1)$$

where $\tilde{c}_n = c_n\, \sinh[2\,H\, z_{0,n}/R]$.

The set $\{J_{0,n}(x)\}_{n=1}^{\infty}$ is a basis of the set of function in the interval $[0,1)$ and they are mutually orthogonal therein with respect to the measure $d\mu(x) = x\,dx$ [2]:

$$\int_0^1 dx\, x\, J_{0,n}(x)\, J_{0,m}(x) = \frac{J_1(z_{0,n})^2}{2}\,\delta_{m,n}$$

The coefficients $\tilde{c}_n$ are therefore found to be the (properly normalized) inner products between the function $f(x) = 1$ and $J_{0,n}(x)$ within $[0,1)$:

$$\tilde{c}_n = \frac{2}{J_1(z_{0,n})^2}\int_0^1 dx\, x\, J_{0,n}(x) = \frac{2}{J_1(z_{0,n})\, z_{0,n}}$$

so that the full solution Z of the Laplace equation is

$$Z(\rho,\, z) = \sum_{n=1}^{\infty} \frac{2}{J_1(z_{0,n})\, z_{0,n}} \frac{\sinh[z_{0,n}\,(z+H)/R]}{\sinh[2H\, z_{0,n}\,/R]}\, J_0(z_{0,n}\rho/R)$$

In the limit $H/R \to \infty$ (infinite cylinder) with finite z, only the first term of the

[2] For $m \neq n$ one can use Gradshteyn–Ryzhik, 6.521, to check the orthogonality condition; for $m = n$ one can integrate twice by parts using $d(x\,J_1(x))/dx = x\,J_0(x)$.

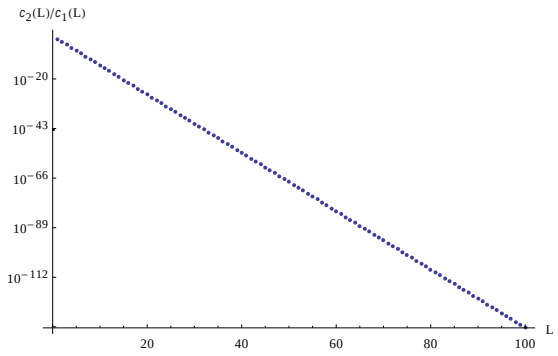


Figure 3.14: The ratio between the coefficients of the first subleading term and the leading one against the length of the cylinder $L = 2H$, in logarithmic scale: the suppression of the subleading terms is exponential in L.

expansion can be retained:

$$Z(\rho,\, z) \propto \exp(z_{0,1}\, z/R)\, J_0(z_{0,1}\, \rho/R)$$

The drift in the effective Langevin dynamics of the conditioned Brownian motion is then

$$\mathbf{u}_*(\rho,\, z) = 2\, D\, \boldsymbol{\nabla} \log Z(\rho,\, z) = 2D\lambda\, \mathbf{e}_z - 2D\lambda\, \frac{J_1(z_{0,1}\rho/R)}{J_0(z_{0,1}\rho/R)}\, \mathbf{e}_\rho,$$

where $\lambda = z_{0,1}/R$.

A.9 Density fluctuations

To study the density of beads along the cylinder, we focus on the evolution of the driven Brownian walker (see Appendix A.7) along the z-axis, described by the stochastic process:

$$\mathrm{d}z_t = 2D\lambda\, \mathrm{d}t + \sqrt{2D}\, \mathrm{d}W_t \tag{A.30}$$

where $\mathrm{d}W_t$ is the standard Wiener process. As described in Appendix A.7 the first term is the drift along the z-axis due to confinement. Consider now the interval $[0,\, \Delta]$ along the z-axis, we define the residence time of the walker therein as

$$\phi_\Delta = \int_0^\infty \mathrm{d}t\, \mathbb{I}_\Delta(z_t)\,, \tag{A.31}$$

where $\mathbb{I}_\Delta$ is the characteristic function of $[0,\, \Delta]$, equal to 1 within the interval and 0 otherwise. In general, ϕ_Δ is a random variable, whose statistics depends on the initial conditions of the process. Its moment generating function is defined as

$$G_\Delta(s,\, z_0) = \left\langle e^{-s\phi_\Delta} \,\middle|\, z_0 \right\rangle \tag{A.32}$$

and satisfies the stationary Feynman–Kac equation

$$2D\lambda \frac{\partial G_\Delta}{\partial z_0} + D \frac{\partial^2 G_\Delta}{\partial z_0^2} = s\, \mathbb{I}_\Delta(z_0)\ . \tag{A.33}$$

In Eq. (A.32), the average is taken with respect to the measure of the paths generated by the dynamics in Eq. (A.30). The drift in Eq. (A.30), that drives the process towards increasing values of z_t, fixes the boundary conditions of $G_\Delta(s,\, z_0)$:

$$\begin{cases} G_\Delta(s,\, z_0) \underset{z_0 \to +\infty}{\longrightarrow} 1\,, & \text{as } \ \phi_\Delta \to 0 \\ G_\Delta(s,\, z_0) \underset{z_0 \to -\infty}{\longrightarrow} \text{const}(s)\,. \end{cases} \tag{A.34}$$

The general solution of Eq. (A.33) then reads

$$G_\Delta(s,\, z_0) = \begin{cases} A_l\, e^{-2\lambda\, z_0} + B_l & \text{for } z_0 < 0 \\ e^{-\lambda z_0}(A_+ e^{\alpha\, x_0} - A_- e^{-\alpha\, z_0}) & \text{for } z_0 \in [0, \Delta] \\ A_r\, e^{-2\lambda\, z_0} + B_r & \text{for } z_0 > \Delta \end{cases}$$

where $\alpha = \sqrt{\lambda^2 + s/D}$ and the A_i and B_i are constants with respect to z_0. The conditions of Eq. (A.34) then set

$$A_l = 0 \qquad \text{and} \qquad B_r = 1\ . \tag{A.35}$$

The four other constants are uniquely determined by imposing continuity and differentiability of G_Δ at $z_0 = 0$ and $z_0 = \Delta$. Note that, for $z_0 < 0$, $G_\Delta(s,\, z_0)$ doesn't depend on z_0, and thus, the statistics of ϕ_Δ are independent of the specific value of z_0.

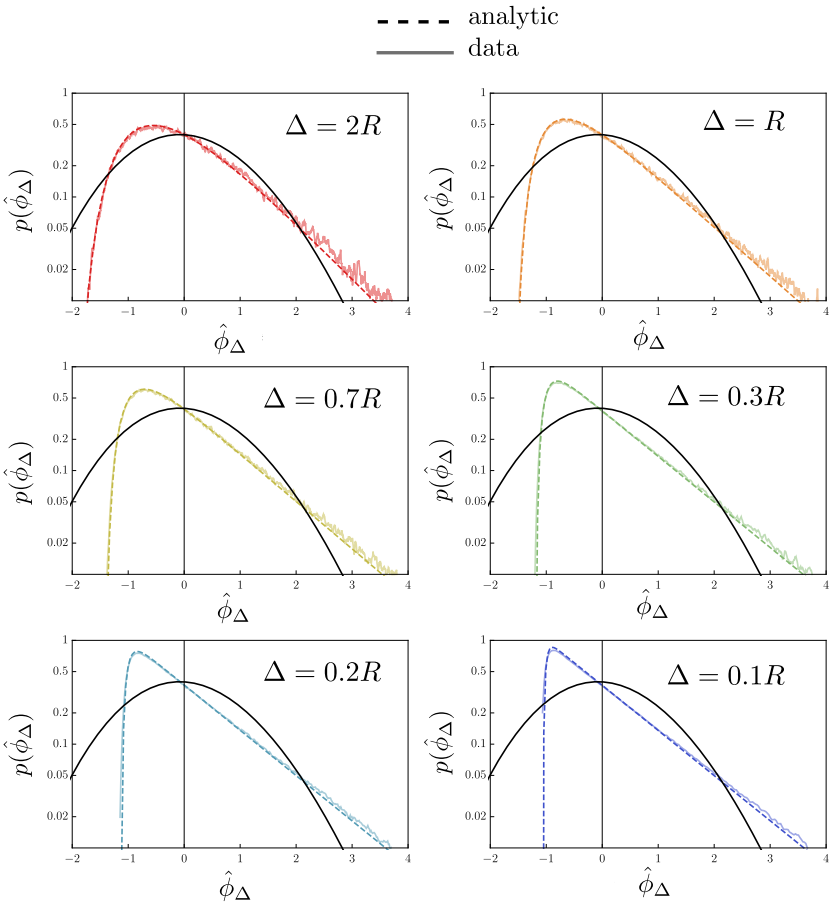


Figure 3.15: Probability density functions of the rescaled residence time $\hat{\phi}_\Delta = (\phi_\Delta - \langle\phi_\Delta\rangle)/\sigma_{\phi_\Delta}$ in the homogeneous case (black curve) and for decreasing values of Δ (colored curves) in the case $\ell_f/R = 4.511 \cdot 10^{-2}$. As discussed in the main text the peak and the tail of each curve highlight inhomogeneities in the system. The right tail indicates the presence of regions with higher density of points and the peak at negative values stems from the presence of regions with lower density of points. The theoretical result (dashed lines) is also shown to be in extremely good agreement with data.

In our estimates of ϕ_Δ, we are interested only in the case $z_0 < 0$, since the initial condition of the process is always to the left of the interval $[0, \Delta]$. Therefore, the moment generating function of ϕ_Δ is given by the amplitude B_l:

$$G_\Delta(s,\, z_0 < 0) = \frac{4\lambda\,\alpha\, e^{\Delta(\alpha+\lambda)}}{(\alpha+\lambda)^2\, e^{2\Delta\alpha} - (\lambda-\alpha)^2}\,, \tag{A.36}$$

where we recall that $\alpha(s) = \sqrt{\lambda^2 + s/D}$. Note that G_Δ can be written in the scaling

form $G_\Delta(s) = \tilde{g}(\Delta^2 s/D,\, \Delta\,\lambda)$, where $\tilde{g}$ is

$$\tilde{g}(u,\, v) = \frac{4v\sqrt{u+v^2}\, e^{\Delta\left(\sqrt{u+v^2}+v\right)}}{\left(\sqrt{u+v^2}+v\right)^2 e^{2\sqrt{u+v^2}} - \left(\lambda - \sqrt{u+v^2}\right)^2}\,.$$

In particular, the diffusion constant D can be absorbed in the scaling variable u. Hence it follows that the probability density of the residence time ϕ_Δ, denoted $F_\Delta(\phi_\Delta)$, is given by the inverse Laplace transform

$$\begin{aligned} F_\Delta(\phi_\Delta) &= \frac{1}{2\pi\mathrm{i}} \int_\gamma \mathrm{d}s\, e^{s\,\phi_\Delta}\, G_\Delta(s)\,, \\ &= \frac{D}{\Delta^2}\, \tilde{f}\left(\frac{D\,\phi_\Delta}{\Delta^2},\, \Delta\lambda\right)\,, \end{aligned} \tag{A.37}$$

where $\tilde{f}$ is the inverse Laplace transform of $\tilde{g}$ with respect to its first variable. The results are shown in Fig. 3.15.

Bibliography

[1] Berg, H. C. *Random Walks in Biology*. Princeton University Press, 1993.

[2] Codling, E. A., Plank, M. J., and Benhamou, S. Random walk models in biology. *Journal of the Royal Society Interface*, 5, 2008.

[3] Frey, E. and Kroy, K. Brownian motion: a paradigm of soft matter and biological physics. *Annalen der Physik*, (1-3), 2005.

[4] Okubo, A. and Levin, S. *Diffusion and Ecological Problems*. Springer, 2nd edition, 2001.

[5] Karatzas, I. and Shreve, S. E. *Methods of mathematical finance*. Number 39 in Applications of mathematics. Springer, New York, NY [u.a.], 1998.

[6] van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*. North-Holland, 1981.

[7] Diao, Y., Ernst, C., Montemayor, A., and Ziegler, U. Generating equilateral random polygons in confinement. *Journal of Physics A: Mathematical and Theoretical*, 44(40):405202, 2011.

[8] Bucklew, J. *Introduction to rare event simulation*. Springer Science, 2004.

[9] Doob, J. L. Conditional Brownian motion and the boundary limits of harmonic functions. *Bulletin de la Société Mathématique de France*, 85:431–458, 1957.

[10] Majumdar, S. N. and Orland, H. Effective Langevin equations for constrained stochastic processes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6), 2015.

[11] Chetrite, R. and Touchette, H. Nonequilibrium Markov processes conditioned on large deviations. *Annales Henri Poincaré*, 16(9):2005–2057, 2015.

[12] Chetrite, R. and Touchette, H. Variational and optimal control representations of conditioned and driven processes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(12), 2015.

[13] Angeletti, F. and Touchette, H. Diffusions conditioned on occupation measures. *Journal of Mathematical Physics*, 57(2), 2016.

[14] Kappen, H. J. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.*, 95:200201, Nov 2005.

[15] Kappen, H. J. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011–P11011, nov 2005.

[16] Todorov, E. Linearly-solvable Markov decision problems. *Advances in Neural Information Processing Systems*, (1):8, 2006.

[17] Todorov, E. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

[18] Tree, D. R., Wang, Y., and Dorfman, K. D. Extension of DNA in a nanochannel as a rod-to-coil transition. *Phys. Rev. Lett.*, 110, 2013.

[19] Odijk, T. The statistics and dynamics of confined or entangled stiff polymers. *Macromolecules*, 16(8):1340–1344, 1983.

[20] Micheletti, C., Marenduzzo, D., and Orlandini, E. Polymers with spatial or topological constraints: Theoretical and computational results. *Physics Reports*, 504(1):1–73, 2011.

Chapter 4

Homing pigeons

A partially observable decision process approach

1 Homing pigeons: the phenomenology

Homing pigeons (*Columba Livia*) have been exploited as message carriers since ancient Egypt. Human interest in homing pigeons started with exploiting their ability to carry messages covering long distances and evolved to a more scientific one with the purpose to understand how they do it. Navigation in birds is a phenomenon that involves many different spatial scales: from migratory processes where birds covering tens of thousands of kilometers [1] to local pinpointing of target locations mostly guided by vision. Homing takes place between these two spatial scales. Aside from being a stage in bird navigation, homing can be operationally defined as the bird ability to find its home when displaced to unfamiliar places. Very well-trained racing pigeons have been reported to fly back home from distances up to a thousand kilometers in homing pigeon races. A record flight for a U.S. Army pigeon is 3700 km in a single flight but single flights of 1600 km were routine.

Apart from these cases, in controlled experiments the spatial scale typically ranges from fairly short distances (around 50 km) to distances that extend up to 400-500 km. Pigeons used in experiments usually are from 4 to 6 months old. They grow in their home site (often called *loft*) and when displaced to unfamiliar distant places they are at their first long-flight experience [2].

The "map-and-compass mechanism" proposed by Kramer in the 1950s considers avian navigation as a two-step process (as in the human case). The *map* is a way to identify the position relative to the goal. A *compass* serves to orient and maintain the correct heading to reach that goal.

Kramer's paradigm allowed to disentangle two different aspects of navigation and to study them separately. In the following we will explain how, over the last decades experimental evidence unraveled the role of the sun and the geomagnetic field as a compass and gave to olfaction a special place in the map step [4].

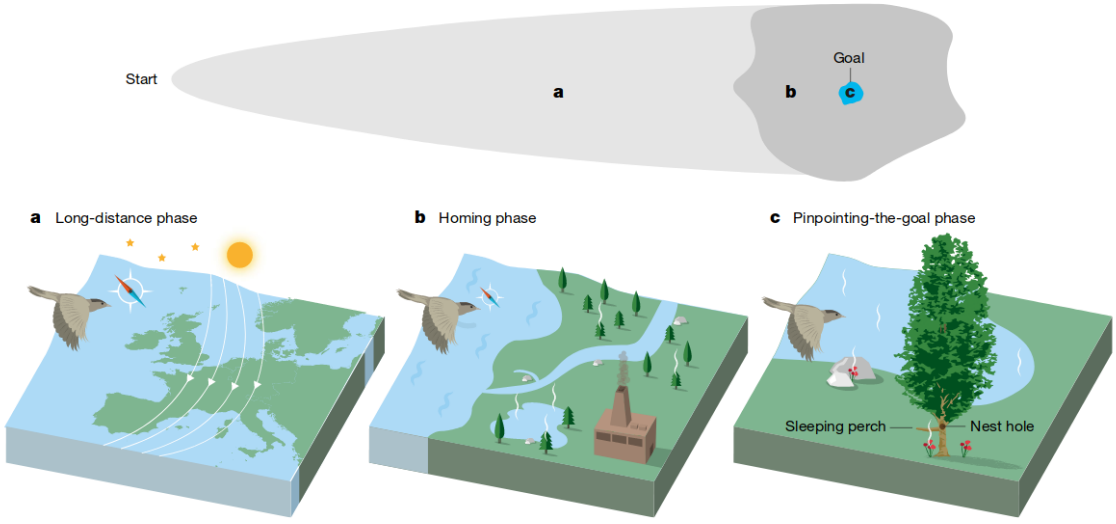


Figure 4.1: **Three different phases of navigation. a.** Mostly guided by celestial cues and by magnetic maps long-distance navigation takes place on the scale of tens of thousands of kilometeres **b.** The homing phase takes place on distances covering hundreds of kilometers **c.** Pinpointing the goal is instead mostly local (up to tens of kilometers) and guided by vision [Picture from Mouritsen (2018)]

1.1 Compass mechanisms: experimental evidences

In the 50s it became clear that pigeons could use the sun to determine compass direction [2]. To keep the compass updated during the apparent movement of the sun requires an internal clock. Coupling the internal clock with the apparent movement of the sun means that at each time of the day they know where the north is with respect to the sun position. With artificial day-night cycles it is possible to shift their internal clock. As a result the angle at which they choose to orient is shifted as well by a quantity predicted by theoretical reasoning (see Fig. 4.2).

Most importantly, the clock-shift experiments show that the sun is a fundamental component in the pigeons' navigation system and that its function is related only to the compass aspect. Even if this mechanism tells the pigeon in which direction it is flying it is not sufficient to reveal if that direction is the correct one. But what do the pigeons do when they are released while the sun is covered by clouds? An experiment by Keeton (1969) puzzled the community presenting results of clock-shifted pigeons that could orient towards home in overcast sky conditions. This made the community think about new experiments concluding that a backup compass system exists. Other results obtained by Keeton brought the scientific community to discover the role of the *geomagnetic field* as an alternative compass mechanism highlighting the importance of the sun as a compass system only in conditions of clear skies. The other effect was to definitely discard the hypothesis that the sun is

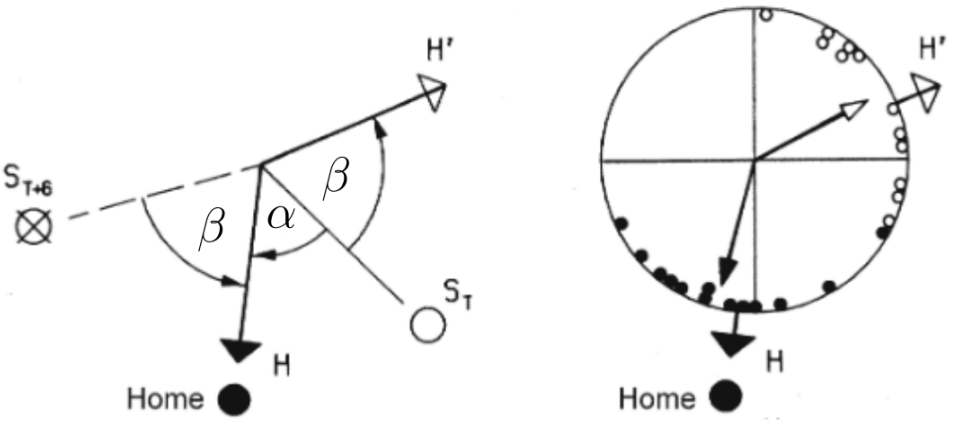


Figure 4.2: **Internal-clock shifting experiments** Let us say our smartphone shows the time T. Control pigeons (•) know that they have to orient at an angle α with respect to the position of sun. If at the same time T of the day we release clock-shifted pigeons (∘) previously exposed to an artificial day-light cycle (e.g. advanced by 6 hours) they will keep an angle β equal to the one they would choose with the sun of $T + 6$h (⊗). The theoretical prediction is that their orientation will be at an angle $\alpha + \beta$ from the home. The left panel shows the theoretical expectation. Right panel: experimental results. `H` and `H'` represent the position of the true home and the theoretical predicted one under time shift. The arrows inside the circle represent the average direction of controls (▶) and operated pigeons (▷). [Adapted from [2]]

used both as a compass and a map.

A series of other experiments accumulated evidences that the earth magnetic field is involved in the compass process when the sun is not available. In fact, Keeton (1971) noticed that experienced pigeons with magnets glued to their head were often showing disorientation when released under total overcast skies, whereas no such disorientation occurred during similar releases under clear skies (see Fig. 4.3). Walcott and Green (1974) equipped the pigeons with Helmholtz coils around the head producing a magnetic field comparable to Earth's magnetic field in two different directions. The results can be summarized as follows: in the presence of the sun both directions of the magnetic field did not produce any alteration on the homing performances but in overcast conditions the opposite directions of the artificial magnetic field caused the pigeons to orient in opposite directions (see fig. 4.4).

Empirical evidence supports the conclusion that pigeons can use the geomagnetic field to determine compass directions, like other birds do. Hierarchically, however, the magnetic compass ranks lower than the sun compass, which is obviously preferred as long as the sun is visible.

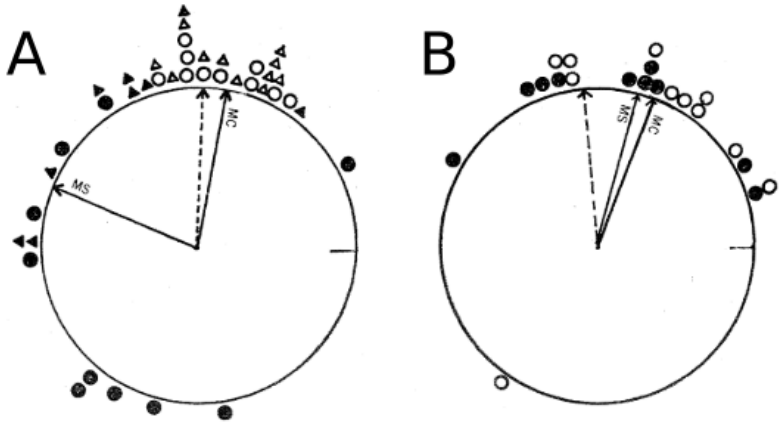


Figure 4.3: **Overcast skies homing. A)** In the sun, the vanishing bearings of the clock-shifted birds were deflected in the direction predicted by the internal clock and sun position coupling. **B)** In overcast sky they were homeward-oriented without being significantly different from the controls. Open symbols (▷ ○) represent the controls in different days. Filled symbols (▶ ●) are the clock-shfted birds. `MC` average direction of controls `MS` average direction of clock-shifted. [Adapted from [5]]

1.2 Olfactory navigation: experimental evidences

Now that the sun and the Earth's magnetic field were identified as two compass mechanisms what was lacking was some environmental cue encoding the map. Between 1950s and 1970s geomagnetism, gravity, celestial bodies, infrasounds and different others stimuli have been proposed to explain how pigeons could obtain positional information [4].

Before 1971 a puzzling issue in the field of bird navigation was the nature of the environmental cues providing pigeons with positional information. Papi et al. (1971) reported that pigeons with sectioned olfactory nerves failed to home. Olfaction appeared for the first time as a possible mechanism for encoding the map in homing pigeons (see Fig. 4.5). Since that discovery, a large empirical evidence in favour of the olfactory hypothesis has been accumulated [2, 10] inducing the conclusion that olfaction is the decisive sense enabling goal-oriented navigation over unfamiliar territories.

The most reliable method of making pigeons definitely anosmic is sectioning the olfactory nerve. If done properly, no nerve reconstitution occurs and the achieved anosmia is permanent. The involvement of the olfactory system in pigeon navigation has not been exclusively demonstrated by lesion experiments. In the brain the piriform cortex receives a large projection from the olfactory bulb and sends projections to numerous other regions of the brain. For this reason it is a good candidate to look for olfactory stimuli processing in the brain. For the first time Patzke et al. (2010)

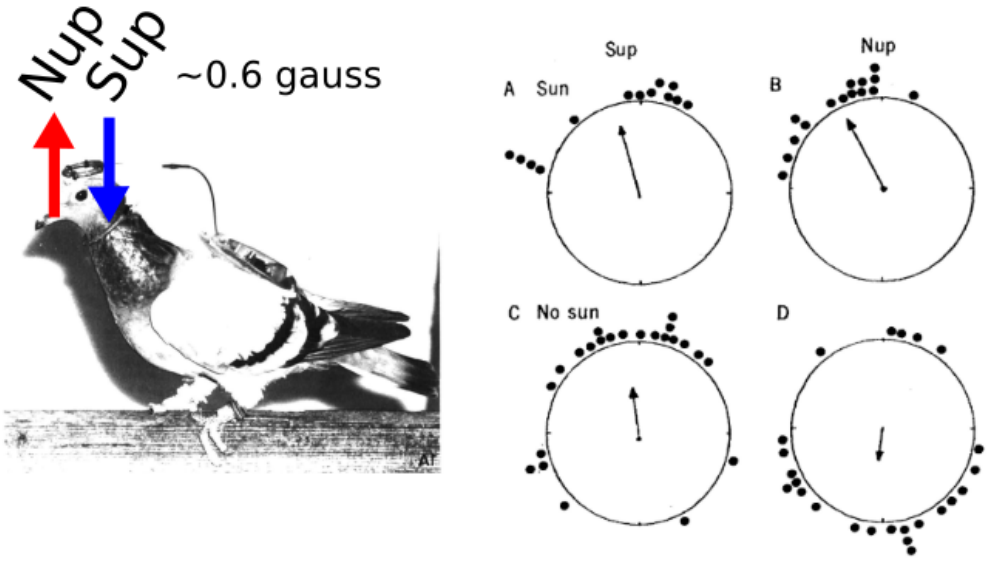


Figure 4.4: **Magnetic fields disturb homing.** Left: A pigeon equipped with Helmholtz coil producing two different magnetic fields (Sup, Nup as in [7]). Adapted from [8]. Right: results from [7]. It is evident that in case of overcast skies, disrupting the sensed magnetic field results in bad homing performances. In both Sup and Nup conditions the pigeons home successfully in sunny conditions (A and B, home is a the top). In case of overcast skies Sup could still orient home but Nup were impaired.

report that the highest number of ZENK-marked cells[1] in the piriform cortex have been found in birds released from unfamiliar places or just exposed to the odors at the unfamiliar site but not released. Instead, pigeons released from a familiar place did not show significant activation.

In the following we would like to focus on a few simple but ingenious experiments shedding light on the important role of olfaction in pigeon navigation.

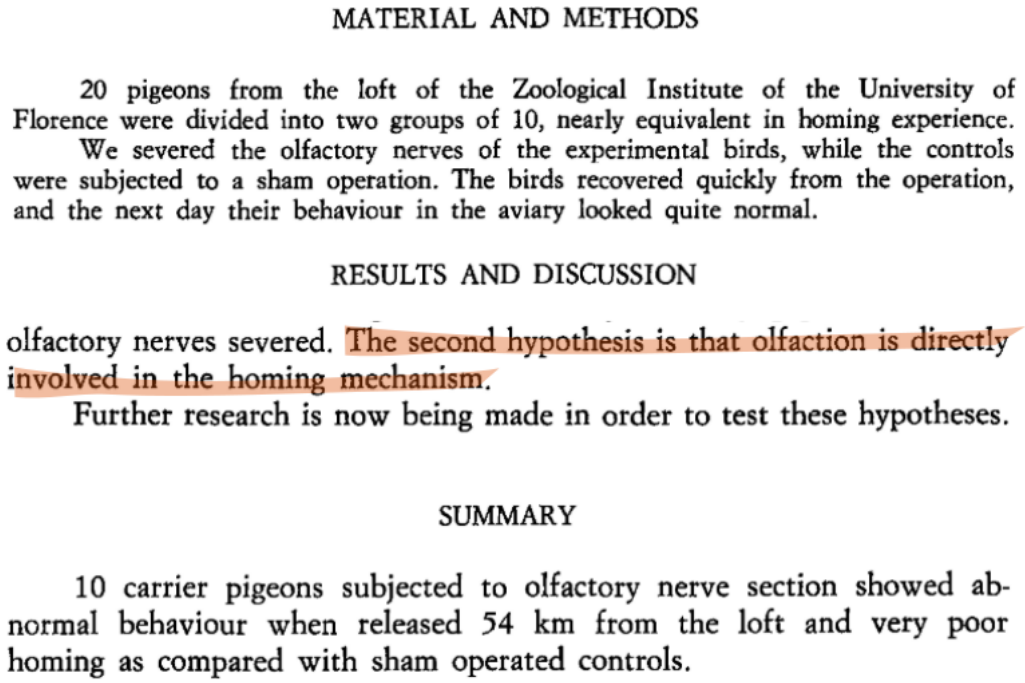
Artificial winds and false release site experiments

A variety of different experimental approaches have revealed two preconditions that must be realized in order to observe homeward-oriented flights of pigeons from unfamiliar distant areas:

1. during the long term stay at the home site the birds must be exposed to winds bringing the surrounding odors

2. at release site the birds must be exposed to environmental air

[1]ZENK is an immediate early gene rapidly expressed in response to external stimuli. An increased expression of the ZENK protein in certain brain regions can be directly linked to neuronal activity



MATERIAL AND METHODS

20 pigeons from the loft of the Zoological Institute of the University of Florence were divided into two groups of 10, nearly equivalent in homing experience.

We severed the olfactory nerves of the experimental birds, while the controls were subjected to a sham operation. The birds recovered quickly from the operation, and the next day their behaviour in the aviary looked quite normal.

RESULTS AND DISCUSSION

olfactory nerves severed. The second hypothesis is that olfaction is directly involved in the homing mechanism.

Further research is now being made in order to test these hypotheses.

SUMMARY

10 carrier pigeons subjected to olfactory nerve section showed abnormal behaviour when released 54 km from the loft and very poor homing as compared with sham operated controls.

Figure 4.5: A snapshot of the paper by Papi et al. (1971) where for the first time the olfactory hypothesis was mentioned as a possible mechanism for homing in pigeons

The first experiment we present modifies the conditions at the home site exposing the birds to artificial winds and odors [12]. In one half of a corridor cage oriented North-South, a group of birds (the "triangles") were exposed to turpentine from North and olive oil from the South. Another group (the "diamonds") in the other half of the cage was exposed to the opposite conditions: olive oil from the South and turpentine from North. After the treatment, both groups were displaced East to an unfamiliar place and exposed to turpentine and olive oil odors, respectively, before the release.

Even if displaced East the "triangles" started flying North when exposed to olive oil odor (the one that was coming from the South at the home site) and South when exposed to turpentine (the odor that was coming from North at the home site). Similar but opposite results were obtained for the "diamonds" demonstrating that pigeons oriented their initial courses according to the associations they had made, at home, between artificial odors and artificial winds. Fig. 4.6 summarizes the results.

Other experiments aimed to manipulate wind directions at the home site, because winds coming from different directions were thought to bring different chemicals from distant sources. In fact, experiments in which the wind direction was artificially deviated at the loft showed a deflection in initial bearing at unfamiliar release site positively correlated with wind deflection at home.4.7 illustrates the idea behind this type of experiments.

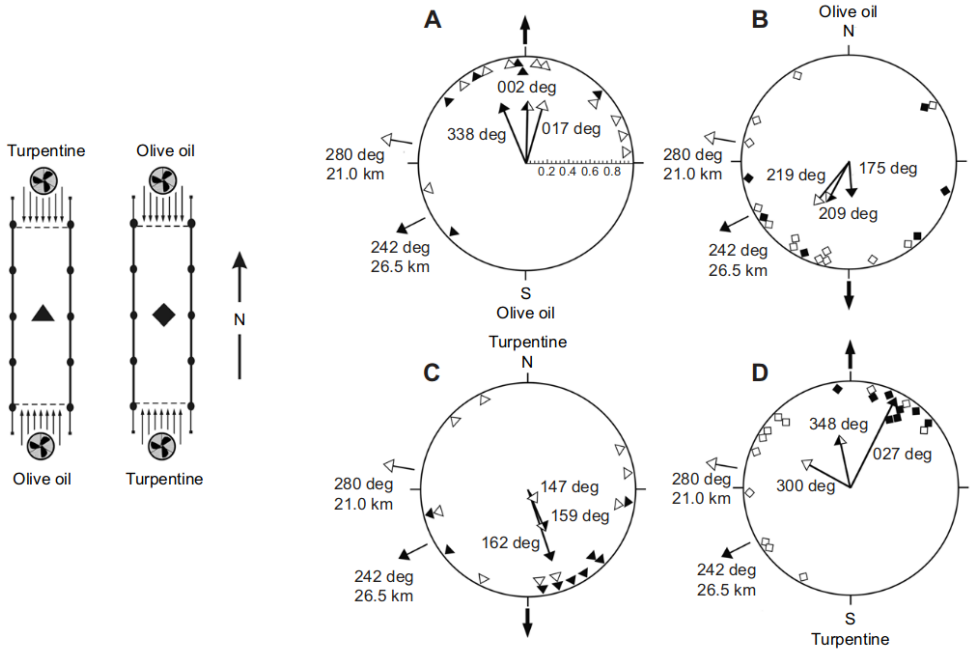


Figure 4.6: **Artificial winds and odors.** Left figure shows two tunnelsm one for the "triangles" pigeons and the other for the "diamonds" pigeons. The tunnels are oriented in the direction N-S as indicated by the arrow. "Triangles" are exposed to turpentine from North and olive oil from South. The opposite happens for "diamonds" pigeons. The right panel shows the results for pigeons displaced West and exposed to the odors experienced at the loft. In summary, even if displaced West they oriented their flight North or South depending on the type of odor they were exposed at the release site.

The second type of experiments we discuss is related to the exposure to odors at the release site. A logical step was to disconnect the site at which pigeons smell local air from the site of release. We can call these experiments false-release-site experiments [2](see Fig. 4.8). The pigeon is transported inside a cage equipped with air filters from the home site to a false release site. At the false release site the air filters are removed and the pigeons are free to smell local odors for some hours. After putting back the air filters they are transported to the true-release-site where they are made anosmic and then released. Many experiments have been conducted (see [2] and references therein) pointing always to the same result: the pigeons treated in this way vanished in directions that were appropriate to the site of smelling but not to the site at which they were actually released suggesting that at the false release site the odor to which they were exposed encoded the direction of home. Moreover, on the same line, other experiments highlight the fact that after sitting for an hour or more in an airtight container ventilated with outside air at the release site, pigeons released under nasal anaesthesia, behave differently depending on the pre-release conditions:

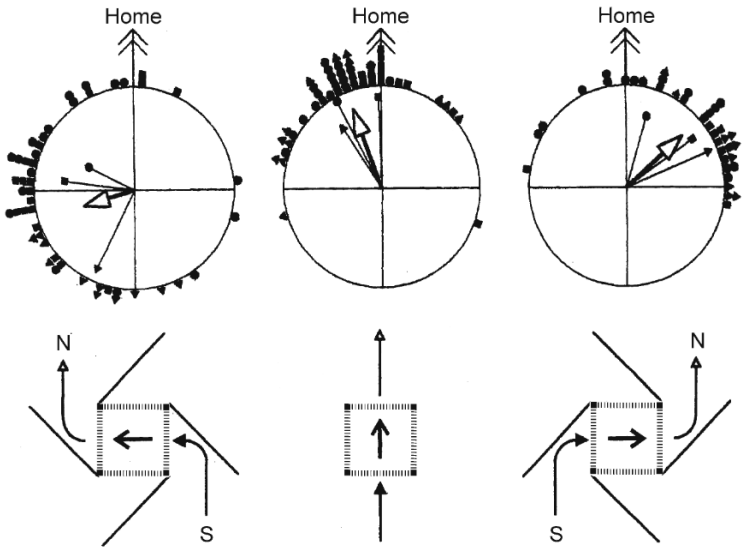


Figure 4.7: **Deflected wind cause rotation of homeward direction.** Bottom panel: using oriented barriers the winds appraoching the home from the soutern site are deflected in such a way that inside the cage the wind is perceived in a rotated fashion. Top panel: the initial homeward direction is correlated to the direction (see [2]).

if they had been allowed to smell natural local air, they depart homeward-oriented; if the air had been filtered removing airborne trace gases, they depart disoriented.

In the last decades the numerous and repeated experiments revealed the nature of physical cues and environmental factors crucially involved in pigeon home-finding process. They can be outlined as follows:

1. a general reference direction can be determined using the sun azimuth or the geomagnetic field with the sun dominating over the geomagnetic system in clear sky days

2. the directions of the winds bringing odors at the home site are crucial to build an angular map of odors with respect to the compass reference

3. atmospheric odors correlated with wind direction at home are used to navigate or at least to orient the flight from unfamiliar places

4. in familiar areas navigation based on olfaction it is not the only source of positional information; instead landmarks also play a role.

1.3 How Does Olfactory Navigation Operate?

This question is still unanswered but Wallraff has drawn on the experimental evidence to propose a possible mechanism that could represent a navigation strategy for the

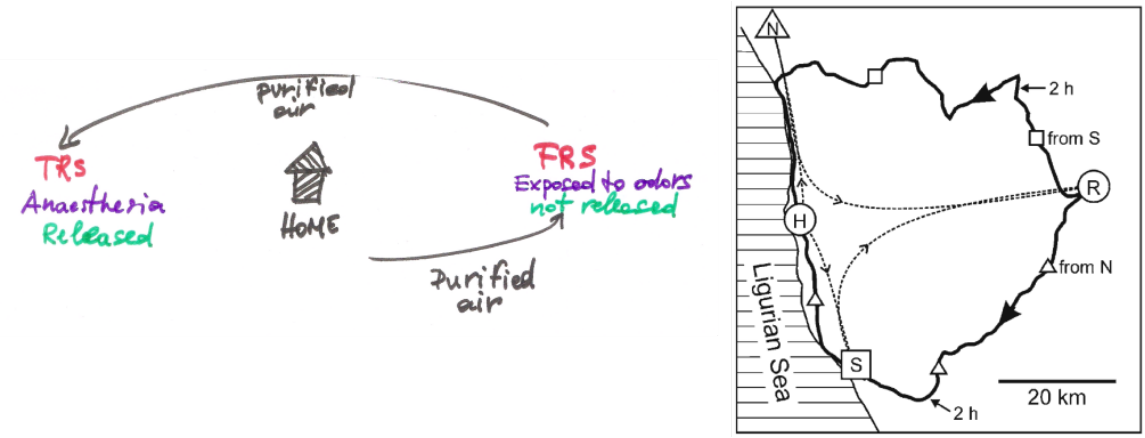


Figure 4.8: **False-release-site experiments.** Left: a scheme of the rationale of the false-release-site experiments. From the home site they are transported in purified air to the false-release-site (FRS) where they are exposed to local air. They are transported again in purified air towards the true-release-site (TRS) where they are made anosmic and then released. Right: Two trajectories from two different groups of pigeons transported from the home site (H) to the false-release-site at N (△) or S (□) and then to the true-release-site R. The triangle pigeon exposed to odors at N headed South and the other familiar with odors in S headed North. The outward path is represented by dotted lines. (see [2] for more details).

pigeons. It consists of two stages, *learning* the atmospheric patterns at the home site and then *homing* using the local knowledge learned at the loft. In the following we introduce with an example learning and homing as depicted by Wallraff [13, 14]. Let us assume that there is a source of a chemical `X` (e.g. a forest) North of the loft. Another source of the chemical `Y` is represented by the sea in the far East.

1. *Learning:* at the home site the wind blowing from North brings on average higher concentration of `X` with respect to other compounds. Winds from East instead correlates more with the compound `Y`. During the period spent at the loft the pigeon learns to correlate wind directions[2] and concentration of odors. In doing so, it builds a sort of angular map for the concentration of each odor. At home the compound `X` will be identified by average concentration $\bar{c}_X$ and direction North and `Y` with $\bar{c}_Y$ and direction East

2. *Homing:* we can now release the pigeon North-West. It will perceive a concentration $c_X > \bar{c}_X$ and $c_Y < \bar{c}_Y$ meaning that it is nearer to the North and farther from the East source of chemical. For this reason it will orient South-East pointing homeward.

[2]The presence of a compass system ensure that the pigeon has a way to define a reference direction.

The idea of navigating using a bi-coordinate system seems a possibility in particular for long-distance navigation based on a geomagnetic field map. Animals familiar with Earth's magnetic field conditions at home may be able to use a grid of isolines related to at least two parameters of the magnetic field to approach the goal area. The precondition that the isolines must be on a regular grid and intersect each other at appreciable angles is not always satisfied ("no-grid zones") and even if the geomagnetic field can be somehow exploited on long-distance navigation (thousands of kilometers) it could not be the case for "local" homing (usually hundreds of kilometers). In any case the geomagnetic grid navigation is still under scrutiny especially for navigation tasks happening in the range of the one we are interested in this work [3, 15].

Moreover due to high variability in space and time induced by turbulence a grid based on atmospheric odors appears at first glance something infeasible but the experimental findings induced the hypothesis that ratios between several chemicals transported by the atmosphere show roughly monotonic spatial gradients over distances of hundreds of kilometres. Wallraff and Andreae (2000) conducted an experiment in Germany to test this hypothesis sampling the air at 96 sites regularly distributed 25 km apart from each other over an area covering a radius of 200 km around Wurzburg (see 4.9 A). In three different summers, 192 air samples were collected, and a statistical analysis of the gas chromatographic measurements on these samples revealed that such gradients in the ratios between a number of omnipresent hydrocarbons do in fact exist. Subsequent theoretical work based on these data [14, 17] show how a pigeon can navigate from site to site (the sampled ones) comparing the concentration of odors at the site with the one at home and transforming this scalar information into a displacement taking into account the angular map of odors built at home. In practice in the *learning* period for each odor `X` the pigeon determines a direction $\hat{g}_X$ from which it perceives the highest concentration for that odor `X`. During *homing* it obtains a flight direction as $\sum_X (c_X - \bar{c}_X)\hat{g}_X$. Trajectories obtained in this way are reported in Fig 4.9 D.

As the authors report this study aims to show that is possible to suggest a navigation system that successfully exploit the information contained in the data of the spatial gradients. To the best of our knowledge this attempt is the only theoretical contribution in terms of algorithms to an open problem but it clearly suffers from two main issues

- in practice the navigation occurs on static landscape and the source of spatial information,even if interpolated, is restricted to a few site separated by 25 km from each other;

- in general a real odor landscape would involve variability in time, as well as in space, over timescale related to the winds' daily variability. It can be that the results obtained by [16] depend too much on the specifics of the particular geographical area.

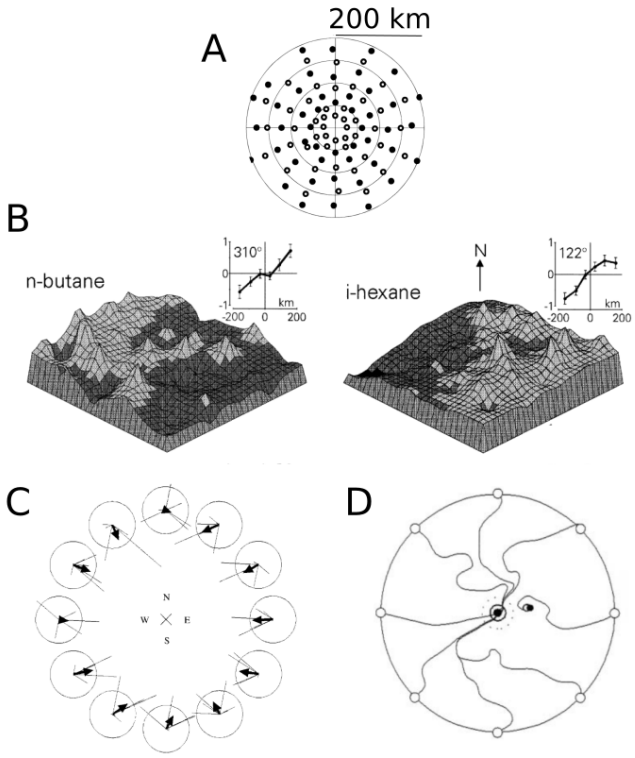


Figure 4.9: **A)** Arrangement of the 192 sites around the pigeon loft (in the center) **B)** two examples of spatial profiles obtained from the air samples and then interpolated to obtain the landscape. The inset shows the spatial trend of each chemical along a given axis (the angle reported in the top left corner) **C)** Computation of the bearing according to six chemical compounds. **D)** Example of trajectories obtained using computed direction as in panel C with 1 km spatial step and interpolated odor landscapes. The central black dot is the home site. Dotted circle represent an absorbing region of radius 10 km. The other black dot represents a stuck trajectory due to zero bearing vector.

In the next section we propose a theoretical framework with which we abstract the process of home finding into a decision making algorithm keeping the fundamental aspects highlighted by the experiments we exposed.

2 Finding home: a partially observable decision process approach

The experiments summarized in the previous section state the following key aspects about homing pigeon:

- they have a compass system to fix a reference direction
- they can use odors to navigate the environment
- the learning phase at home is crucial for them to understand two quantities: the average concentration of a given odor and its relevant direction bringing on average higher concentration of that volatile compound
- manipulation of local atmospheric elements at the home site (e.g. deflecting the winds or fooling the pigeon about the release site) results in disorientation at the release site

Note that a necessary comment must be made about the last point. When pigeons are exposed to rotated winds at home and then released at the unfamiliar site they show initial disorientation. Even if vanishing bearing results (see Fig. 4.7) tell us that they orient in the wrong direction with respect to home, looking at the trajectories show that pigeons can correct their initial *belief* and eventually get successfully to the loft. The same can be said about false-release-site experiments. We will comment later on this aspect in relation to what we propose in our model.

2.1 Analytical solution for the one dimensional case

Now we would like to cast the home finding problem by pigeons in a dynamic odor landscape into a Partially Observable Markov Decision Process (POMDP) as introduced in Sec. 2.3. For the sake of simplicity we will start with a one dimensional homing task in a single odor landscape, show its POMDP structure and then extend the framework to the general case of two dimensional homing in the presence of many odors. We consider an odor landscape represented by a large scale increasing gradient g on which we superimpose a Gaussian noise. This means that at each location x the pigeon will perceive an odor

$$c(x) = gx + \eta$$

with $\eta \sim N(0, \sigma_c^2)$ a Gaussian variable with zero mean and variance σ_c^2. Before introducing the quantities to define the POMDP framework we define the task. We fix the home site at $x = 0$ and without loss of generality the average concentration at home will be $\bar{c} = 0$. Such an odor landscape can be considered a simplified situation in which there is a source of odor far away from home in the positive axis (on the right).

In this setting finding the home coordinate is equivalent to finding the zero of the function $f(x) = gx$ by means of observations c that are independent and identically distributed Gaussian variables representing the noisy odor signal the pigeon perceives

during navigation[3]. Following the olfactory hypothesis (see [4, 10]) the general idea is that the odor perceived at a given location encode somehow for the spatial coordinate. The minimum number of odors needed will be related to the dimension of the domain. In a one-dimensional navigation task one odor is sufficient but if we want to navigate a two-dimensional environment at least two odors are necessary.

As we already said at the home site the pigeon can define a direction from which the average concentration over time is higher. In this synthetic odor landscape we can identify such direction with the positive direction of the real axis. To make a parallel with what we previously said imagine to have a "one dimensional wind" blowing on the real axis in both directions according to some periodicity. On average this wind will bring higher concentration from the right then from the left. Anyway the information about this direction does not bring any cue about the strength of g and we consider g an unknown of the problem.

We can now define the POMDP and solve for the optimal Bellman equation associated to it. The plan is the following:

1. we define the POMDP (see the box at the end of Sec. 2.3): the model of the environment, the likelihood, the belief and the reward function
2. we write and solve the optimal Bellman equation that will give us the optimal policy for the problem

We define the state $\bar{s}$ of the environment as the true value of the odor at location x and the gradient itself

$$\bar{s} = (s \equiv gx, g)$$

The model of the environment will be defined by

$$p(s', g'|s, g, a) = \delta(s' - s - ga)\delta(g' - g) \tag{1}$$

where $\delta(x)$ is the Dirac delta. Here a represents the action. In this case it can take just two values: "left" and "right". For the moment we do not fix its magnitude. The other object to identify is the likelihood of the observation. Due to the structure of the odor landscape, sitting at a location x with odor concentration s the likelihood to get an observation c is

$$\ell(c|s) = N(y|s, \sigma_c^2) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[\frac{(c-s)^2}{2\sigma_y^2}\right] \tag{2}$$

Finally we define the reward function to be dependent only on s:

$$r(s) = -s^2$$

[3] we can think to define the functiion $f(x)$ as an average of the observations c conditioned to the point x: $f(x) = \mathbb{E}(c|x)$

In this way the closer the agent is to home the higher is the reward. In this context we can interpret the homing behavior as the tendency of the pigeon to be closer and closer to the loft characterized by a given averager odor concentration [4].

According to the Bayesian update eq.(2.11) of Chapter 2 the belief to be in a given state s will evolve according to the model of the environment (1) and the likelihood (2). Choosing a Gaussian belief means that the prior and the posterior will be related by simple relations and moreover the belief will remain Gaussian. For this reason we can take

$$b(\overline{s}) = \frac{\sqrt{\det M}}{2\pi} \exp[-\frac{1}{2}(\overline{s} - \overline{\mu})^T M (\overline{s} - \overline{\mu})]$$

where M is the inverse of the 2×2 covariance matrix with entries

$$M = \begin{bmatrix} M^{ss} & M^{sg} \\ M^{gs} & M^{gg} \end{bmatrix} \tag{3}$$

where $M^{sg} = M^{gs}$ and $\overline{\mu} = (\mu_s, \mu_g)$ are the estimates for the odor level and the value of the gradient respectively.

In this context the space of states $\overline{s}$ and observations c is continuous. This means that the $\sum_s$ and $\sum_o$ in the expressions discussed in Sec. 2.3 become integrals. Thus

$$b'(s', g|y, a) = \frac{\int ds b(s, g) \ell(c|s') p(s'|s, g, a)}{\int ds' dg b(s, g) \ell(c|s') p(s'|s, g, a)} = \frac{b(s' - ga, g) f(y|s')}{\int ds' dg b(s' - ga, g) \ell(c|s')} \tag{4}$$

that is still a Gaussian with covariance matrix M'

$$\begin{aligned} M'^{ss} &= M^{ss} + \sigma_y^{-2} \\ M'^{sg} &= M^{sg} - a M^{ss} \\ M'^{gg} &= M^{gg} + a^2 M^{ss} - 2a M^{sg} \end{aligned} \tag{5}$$

These expressions come from the comparison of the second order terms in s' and g at the exponent of the numerator in eq. (4) (see Appendix 3).

Instead the average of the new belief will be defined by

$$\begin{aligned} \mu'^s &= \mu^s + a\mu^g + \alpha\big[c - (\mu^s + a\mu^g)\big] \\ \mu'^g &= \mu^g - \beta\big[c - (\mu^s + a\mu^g)\big] \end{aligned} \tag{6}$$

with α and β the two learning rates

$$\alpha = \frac{1}{\sigma_c^2}(M'^{-1})^{ss} \ , \quad \beta = \frac{1}{\sigma_c^2}(M'^{-1})^{sg} \tag{7}$$

[4] We set $s = 0$ at the loft but in general the reward function can be redefined to be $r(s) = -(s - s_{Home})^2$

According to these equations, an agent can start with an initial belief M_0,$\overline{\mu}_0$ and use the observations c to update the estimates for the covariance matrix M and the averages $\overline{\mu}$ thus refining the belief.

The next step is to find a way to map the belief into optimal actions. This means we have to solve the optimal Bellman equation associated to this problem.

But before we would like to cast the equations above in a more simplified form. Starting from $M_0 = 0$ (a completely flat belief) we can rewrite the recursive relation (5) as

$$\begin{aligned} M_t^{ss} &= \frac{t}{\sigma_c^2} \\ M_t^{sg} &= -\frac{\Delta_t}{\sigma_c^2} \\ M_t^{gg} &= \frac{1}{t}\left[\sum_{\tau=2}^{t-1} \frac{\Delta_\tau^2}{\tau(\tau-1)}\right] + \frac{1}{\sigma_c^2(t-1)}\Delta_t^2 \end{aligned} \tag{8}$$

with

$$\begin{aligned} \Delta_t &= \sum_{\tau=0}^{t-1} \tau a_\tau = \Delta_{t-1} + (t-1)a_{t-1} \\ H_t &= \sum_{\tau=2}^{t} \frac{1}{\tau(\tau-1)}\Delta_\tau^2 = H_{t-1} + \frac{\Delta_t^2}{t(t-1)} \end{aligned}$$

Using these relations we can rewrite the learning rates (B.28) and finally the equations that learn μ_s and μ_g take the form

$$\begin{aligned} \mu_{t+1}^s &= (\mu_t^s + a\mu_t^g) + \overbrace{\frac{1}{t+1}\left(1 + \frac{\Delta_{t+1}^2 H_{t+1}^{-1}}{t+1}\right)}^{\alpha_{t+1}}(y - (\mu_t^s + a\mu_t^g)) \\ \mu_{t+1}^g &= \mu_t^g - \underbrace{\frac{\Delta_{t+1}H_{t+1}^{-1}}{t+1}}_{\beta_{t+1}}(y - (\mu_t^s + a\mu_t^g)) \end{aligned} \tag{9}$$

We notice that the learning phase will consist to estimate at home the local value of the average concentration $\overline{c}$ and the variance σ_c^2 of the odors. During the homing phase to update the values of μ_s and μ_g is just sufficient to know the expression of H^{-1} (H has a simple iterative form) and Δ that can be updated on the fly.

In the general scheme of the POMDP the set of equations (8) and (9) represent the *state estimator* part of the problem. We now have to look for the optimal policy.

Since the value function depends on the belief in this particular case it will be a function of $\overline{\mu}$ and M. Since the reward does not depend on the action, the Bellman equation in the box at the end of Sec. 2.3 can be rewritten as follows

$$V(\overline{\mu}, M) = \int dsr(s)b(s) + \gamma \max_a \left[\int dc\ell(c|a)V(\overline{\mu}', M') \right] \tag{10}$$

The expected reward takes the form

$$\overline{r} = \int dsr(s)b(s) = -\left(\mu_s^2 + M_{ss}^{-1}\right)$$

We still have to define the nature of the action a. In this context we think it is natural to define a as a discrete step size jump in the only two possible directions thus $a = \pm\delta$. The expression for the expected reward gives us a hint on a possible form for the value function $V(\overline{\mu}, M)$. In fact, if we make the ansatz

$$V(\overline{\mu}, M) = -A(\mu^s\mu^s + M_{ss}^{-1}) - B(\mu_s\mu_g + M_{sg}^{-1}) - C(\mu_g^2 + M_{gg}^{-1})$$

we can solve the Bellman equation provided that we choose

$$A = \frac{1}{1-\gamma} \quad B = \frac{2\gamma a}{(1-\gamma)^2} \quad C = \frac{a^2\gamma(1+\gamma)}{(1-\gamma)^3}. \tag{11}$$

The optimal value takes the form

$$V^*(\overline{\mu}, M) = -\frac{1}{1-\gamma}(\mu^s\mu^s + M_{ss}^{-1}) - \frac{2a}{(1-\gamma)^2}(\mu_s\mu_g + M_{sg}^{-1}) - a^2\frac{1+\gamma}{(1-\gamma)^3}(\mu_g^2 + M_{gg}^{-1}) \tag{12}$$

and finally the optimal policy is defined by

$$a^* = \text{argmax}_a\, V \rightarrow a^* = -\delta\, \text{sign}\left(\mu_s\mu_g + M_{sg}^{-1}\right). \tag{13}$$

Notice that without the learning part expressed by eqs.(9) and perfectly knowing the value of the gradient this policy corresponds to a completely reactive strategy that uses only the noisy odor concentration to select the action. We will comment on this in the following section.

2.2 Preliminary results in one dimension

In the light of this result we would like to interpret and discuss the paradigm summarized at the beginning of Sec. 1.3 (see also Fig.1 of [10]). At the home site the agent is exposed to the wind-borne odor. On average it will perceive higher concentration from the right (direction $+\hat{x}$) and will fix the gradient to be positive. It also learns the average concentration $\overline{c}$ describing the odor at the home site. According to Wallraff's algorithm [13, 18], when displaced far away from the loft, the bird compares the odor concentration at the new location with $\overline{c}$: if it is higher it goes left, if it is lower it goes right. This reactive strategy can be translated into the update rule

$$x_{t+1} = x_t - \hat{g}\,(c - \overline{c}) \tag{14}$$

where $\hat{g}$ is the direction of the gradient and plays a crucial role (in this case $\hat{g} = \pm\hat{x}$). If at home we reverse the wind the agent learns the wrong $\hat{g}$ and the update rule will always fail to bring the agent close to home (see Fig. 4.10 A). As expected all the trajectories starting from both positions far in the left or in the right with respect to home systematically diverge away from the target, defined as a region of width Δ. This is not the case for the POMDP model. As it is shown in Fig. 4.10 B all trajectories starting with the wrong value of the gradient are able. At the beginning the agents keep the same wrong direction as in Wallraff and this effect as expected is stronger the stronger is the belief that the gradient has that negative value. In practice the two algorithms behave the same in the limit $M^0_{gg} \to \infty$. Using the observations of the odor concentration the agent is able to learn both the magnitude and the direction of the gradient, correcting the initial trajectory (see Fig. 4.10 C,D respectively).

Another aspect that cannot be embedded in the update rule (14) is related to the outcomes of the false-release-site experiments. In practice, the agent can gain information about the local odor at the place where is transported but not released. It means that its belief will be peaked around that value of concentration when released from a different place.

2.3 Analytical solution for the multiple odors case

Navigation is obviously a task that take place in space and not in a line. Also the atmosphere typically presents multiple odor sources. The model we presented in the one-dimensional case has then to be extended to describe a situation with a generic number n of odors in d spatial dimensions. We will follow the same steps we presented for the one-dimensional case. Here the calculation are just more convoluted but the idea behind it is exactly the same. In this case we would like to find the zero of the multidimensional function $F(x) = gx$ where $x \in \mathbb{R}^d$, $F \in \mathbb{R}^n$ and g is a $n \times d$ matrix.

The states, hidden to the agent, are defined to be the value of the function $F = gx \equiv s$ and the gradient itself

$$\overline{s} = (s, g) \quad \overline{s} \in \mathbb{R}^{n(d+1)} \tag{15}$$

We define the model of the environment exactly as in the one-dimensional case

$$p(s', g'|s, g, a) = \delta(s' - s - ga)\delta(g' - g). \tag{16}$$

The likelihood $\ell(c|s)$ $N(s, \sigma_c^2)$ of the one dimensional case will be replaced by the multivariate Gaussian

$$\ell(c|s) = \frac{\sqrt{\det K}}{(2\pi)^{n/2}} \exp[-\frac{1}{2}(c - s)^T K(c - s)] \tag{17}$$

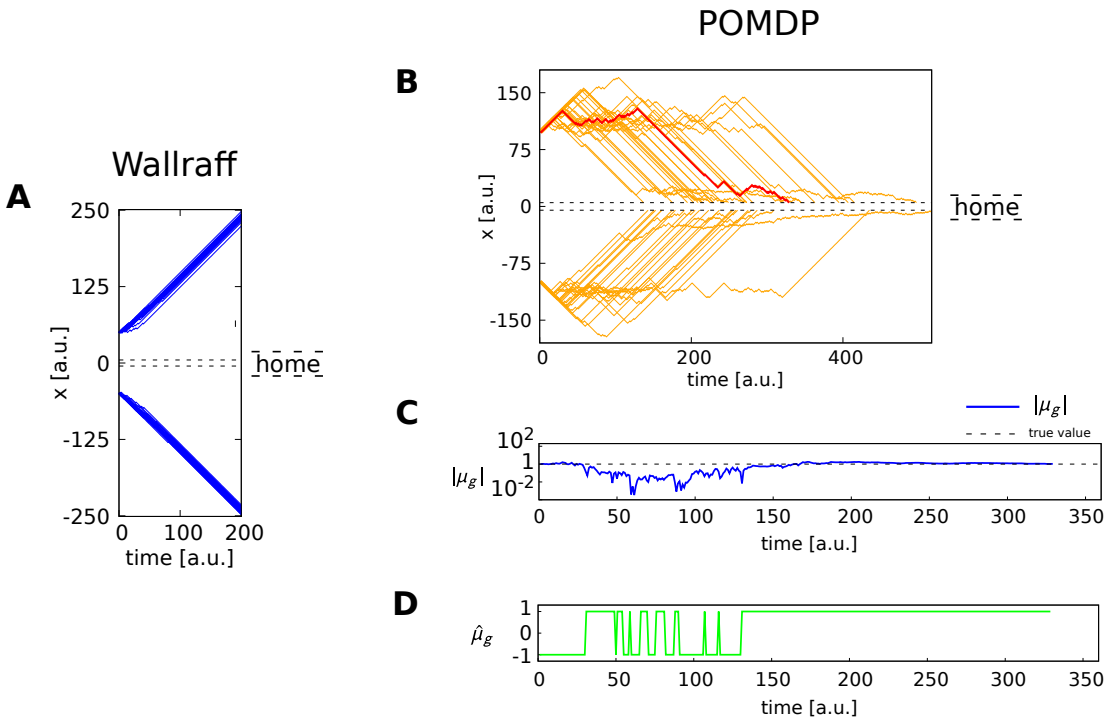


Figure 4.10: **Synthetic wind deflection experiments. A)** This plot shows 50 trajectories starting from $x_0 = \pm 50$ with $\hat{g} = -\hat{x}$. They all diverge from home (the region between dashed lines $\Delta = 5$) **B)** the figure shows 50 trajectories starting from $x_0 = \pm 100$ with the wrong initial gradient $\mu_0^g = -\hat{x}$ and $M_{gg} = 0.25$. Highlighted in red a trajectory for which the learning of the magnitude of the gradient and its direction is shown (panel **C** and **D** respectively). $\sigma_y = 100$

where K^{-1} is the $n \times n$ covariance matrix for the observations (in the one dimensional case we had only σ_c^2 quantifying the noise on the odor concentration). The belief is the Gaussian

$$b(\overline{s}|\overline{\mu}, M) = \frac{\sqrt{\det M}}{(2\pi)^{n(d+1)/2}} \exp[-\frac{1}{2}(\overline{s} - \overline{\mu})^T M (\overline{s} - \overline{\mu}) \tag{18}$$

where $\overline{\mu} = (\mu_s, \mu_g)$ is a $n(d+1)$ vector wtih the first n components given by the vector μ_s of the mean odor concentration and the other nd components by the mean gradient μ_g. Note that the $n(d+1) \times n(d+1)$ matrix M is a block matrix defined by

$$M^{-1} = \begin{pmatrix} M_{ss}^{-1} & M_{sg}^{-1} \\ (M_{sg}^{-1})^T & M_{gg}^{-1} \end{pmatrix} \tag{19}$$

and each block inverse in general is not the inverse of the block (for example $M_{ss}^{-1} \neq (M_{ss})^{-1}$).

Now we have all the ingredients to compute the update rules for the parameters defining the belief: the covariance matrix M and the mean μ. We recall the belief update rule to be

$$b'(\overline{s}'|c,a) = \frac{\ell(c|s') \int d\overline{s}\ b(\overline{s})\ p(\overline{s}'|\overline{s},a)}{\underbrace{\int d\overline{s}' d\overline{s}\ \ell(c|s')\ b(\overline{s})\ p(\overline{s}'|\overline{s},a)}_{\ell(c|a)}} \tag{20}$$

We first focus on the numerator and then on the denominator. Recall that the denominator $\ell(y|a)$ is also called the *evidence* and is of fundamental importance to write down the Bellmann equation associated to the problem. Exploiting the presence of the $\delta(\overline{s}' - s)$ in the model of the environment we integrate over the variable g and s obtaining

$$b'(\overline{s}'|c,a) = \frac{b(s' - ga, g)\ell(c|s')}{\ell(c|a)} \tag{21}$$

The next step is to obtain the explicit update rules for the two variables defining the belief: the covariance matrix M and the average $\overline{\mu}$. We skip here all the calculations (see Appendix **??**) that are based on the fact that the numerator is a product of two multivariate Gaussians and therefore the new belief b' will be a Gaussian. To find the new covariance matrix M' of b' we can collect the terms of quadratic order in s' and g giving

$$\begin{aligned} M'_{ss} &= M_{ss} + K \\ M'_{sg} &= M_{sg} - M_{ss} \otimes a^T \\ M'_{gg} &= M_{gg} - M_{sg} \otimes a - M_{gs} \otimes a^T + M_{ss} \otimes (aa^T) \end{aligned} \tag{22}$$

where a is the action represented by a d-dimensional vector and $\otimes$ represents the Kronecker product.

If instead we collect the terms linear in s' and g we obtain

$$\begin{aligned} M'_{ss}\mu'_s + M'_{sg}\mu'_g &= M_{ss}\mu_s + M_{sg}\mu_g + Kc \\ M'_{gs}\mu'_s + M'_{gg}\mu'_g &= M_{gs}\mu_s + M_{gg}\mu_g - (M_{ss} \otimes a)\mu_s - (M_{sg} \otimes a)\mu_g \end{aligned}$$

The solution for μ'_s and μ'_g is thus

$$\begin{aligned} \mu'_s &= \mu_s + \mu_g a + M_{ss}^{-1} K\left[c - (\mu_s + \mu_g a)\right] \\ \mu'_g &= \mu_g + M_{gs}'^{-1} K\left[c - (\mu_s + \mu_g a)\right] \end{aligned} \tag{23}$$

We can now write the Bellman equation and find its solution. In the one dimensional case a hint to write a tentative form for the value function came from the form we took for the reward function. Also in this case we will take a reward

$$r(s) = -s^T s \to \overline{r} = \int ds\, r(s)\, b(s,g) = -\mu_s^T \mu_s - \operatorname{tr} M_{ss}^{-1} \tag{24}$$

As before the reward for the agent will be higher the closer it is to the reference level (here taken to be zero at the home site). The optimal Bellman equation for the scalar value function $V(b) \equiv V(\overline{\mu}, M)$ is

$$V(\overline{\mu}, M) = \overline{r} + \gamma \max_a \left[\int dc\, \ell(c|a)\, V(\overline{\mu}', M') \right] \tag{25}$$

To gain insight on the shape of V in terms of M and $\overline{\mu}$, without knowing the exact expression for $\ell(c|a)$ we can exploit the fact that (see eq. 21)

$$\ell(c|a) b'(\overline{s}'|c,a) = b(s' - ga, g) \ell(c|s') \tag{26}$$

and try to compute for example the expected value over all the observations c of the reward

$$\begin{aligned}
\int dc\, \ell(c|a)(\mu_s'^T \mu_s' + \operatorname{tr} M_{ss}'^{-1}) &= \int dc\, \ell(c|a) \int ds' dg\, s'^T s'\, b'(s', g|c, a) \quad \text{use (26)} \\
&= \int dc\, ds' dg s^T s' \ell(c|s') b(s' - ga, g) \\
&= \int ds'\, dg\, s'^T s' b(s' - ga, g) \\
&= \int ds\, dg\, (s^T s + 2 s^T g a + a^T g^T g a) b(s, g) \\
&= \underbrace{\mu_s^T \mu_s + \operatorname{tr} M_{ss}^{-1}}_{\overline{r}} + \\
&\quad + \mu_s^T \mu_g a + (\mu_s^T \mu_g a)^T + 2 \operatorname{tr}(M_{sg}^{-1} a) + \\
&\quad + (\mu_g a)^T \mu_g a + \operatorname{tr}(a^T M_{gg}^{-1} a)
\end{aligned} \tag{27}$$

In this expression we recognize the reward (24). This result gives us a hint on how to build the value function. In Appendix 3 we show that the form

$$\begin{aligned}
V(\overline{\mu}, M) = -\Big[& A(\mu_s^T \mu_s + \operatorname{tr} M_{ss}^{-1}) + \\
& + B(\mu_s^T \mu_g a + \operatorname{tr}(M_{sg}^{-1} a)) \\
& + C(a^T \mu_g^T \mu_g a + \operatorname{tr}(a^T M_{gg}^{-1} a)) \Big]
\end{aligned} \tag{28}$$

solves the Belmann equation provided that

$$A = \frac{1}{1-\gamma} \quad B = \frac{2\gamma}{(1-\gamma)^2} \quad C = \frac{\gamma(1+\gamma)}{(1-\gamma)^3}$$

Among a discrete set of predefined actions the optimal one will be defined by

$$\begin{aligned} a^* &= \text{argmax}_a \int dc\, \ell(c|a) V(\overline{\mu}', M') \\ &= \arg\min_a \left[\mu_s^T \mu_g a + \text{tr}(M_{sg}^{-1} a) + \frac{1+\gamma}{2(1-\gamma)} \left(a^T \mu_g^T \mu_g a + \text{tr}(a^T M_{gg}^{-1} a) \right) \right] \end{aligned} \tag{29}$$

Summarizing, Eqs.(22) and (23) represents the analogous of Eqs. (5) and (6). Obviously, if we consider $d = 1$ and $n = 1$ the general expressions (22) , (23) and (29) reduces to the ones we found in the one dimensional case with the substitution $K \to \sigma_c^{-2}$. In fact, in the action selection rule (29) the second term in the sum will give a constant contribution if the action has only two possible values $a = \pm\delta$ and the $\text{tr}(M_{sg}^{-1} a)$ will be trivially equal to $\pm\delta M_{sg}^{-1}$.

Moreover it is not surprising that the algorithm can be rewritten in the same recursive form already seen for the one dimensional case

$$\begin{aligned} \mu_s^{t+1} &= \mu_s^t + \mu_g^t a_t + \alpha_{t+1} \left[c_t - (\mu_s^t + \mu_g^t a_t) \right] \\ \mu_g^{t+1} &= \mu_g^t + \left[c_t - (\mu_s^t + \mu_g^t a_t) \right] \otimes \beta_{t+1}^T \end{aligned} \tag{30}$$

where α_t and β_t are respectively a scalar and $d-$dimensional vector representing the learning rates

$$\alpha_t = \frac{1}{t} \left(1 + \frac{1}{t} \Delta_t^T H_t^{-1} \Delta_t \right)$$

and β_t is the $d-$dimensional vector

$$\beta_t = \frac{1}{t} H_t^{-1} \Delta_t$$

with the $d \times d$ matrix H_t and the $d-$dimensional vector Δ_t having the expressions

$$\begin{aligned} \Delta_t &= \sum_{\tau=0}^{t-1} \tau a_\tau = \Delta_{t-1} + (t-1) a_{t-1} \\ H_t &= \sum_{\tau=2}^{t} \frac{1}{\tau(\tau-1)} \Delta_\tau \otimes \Delta_\tau^T = H_{t-1} + \frac{1}{t(t-1)} \Delta_t \otimes \Delta_t^T \end{aligned}$$

Thus also in this case the agent needs to keep only track of Δ_t and H_t to update the belief and learn the covariance matrix K and the vector of concentration $\overline{c}$ at home.

The algorithm proposed by Wallraff can be extended to the multiple odors case and takes the form

$$x_{t+1} = x_t + \sum_i \hat{g}_i \left(c_i - \overline{c}_i \right)$$

where the sum is done over all the odors. As reported also by Wallraff this algorithm can become inefficient when that sum comes close to zero giving a null bearing direction. An example of an unsuccessful trajectory is reported in Fig. 4.9.

The theoretical result we obtained and the preliminary evidences in the one dimensional case open different directions to investigate. We will discuss them in the following.

3 Discussion

Navigation based on noisy odor cues has been algorithmically approached in different situation from the one we presented here [19, 20]. The situation of homing birds cannot be described for example as the case of the moths searching for mates exploiting a pheromone signal emitted upwind. In our case the target at which a pigeon is directed does not emit any signal. Moreover pigeons home equally well with both head and tail winds. Even if we cannot consider the home site as a source emitting odors, it is however characterized by a given concentration of chemicals. To the best of our knowledge regarding the particular case of homing pigeons, Wallraff [13, 18] was the only one to propose an algorithm based on experimental evidences. This algorithm can be translated into the update rule (14). However, this approach is not completely consistent with the experimental evidence. Odors that are locally manipulated induce the pigeons to lose the initial orientation towards home, but eventually they can find their way home by correcting their route. Wallraff mentions that a completely reactive strategy such as the one produced by that algorithm could result in a "quite chaotic zig-zag course" and he proposes a way to correct this behavior taking into account the trajectory flown by the bird up to a given point to update its new flying direction. In Appendix 3 we show how the history of observations can be taken efficiently into account at fixed gradient. In any case, the crucial fact here is that once the pigeon learns the wrong cues regarding the gradients at home it can never correct them using observations sensed along the trajectory. In our work instead the "previous history" of the bird's trajectory is taken into account in the space of odor and gradient and their estimates can be corrected exploiting new observations. We solved an optimization problem in a very particular scenario where the noisy measurements are Gaussian-distributed and completely uncorrelated. The preliminary insights coming from the very simplified case of a one dimensional navigation task open different directions we would like to follow. A natural one we are already investigating is of course to study the properties of the algorithm in the two dimensional case. A more intriguing one is to understand what are the performances in a more realistic environment where the dynamics of the odors is driven by turbulence. We will start to simulate the turbulent dynamics of two passive scalars representing the odors (see Fig. 4.11 A). The odor landscape will present two profiles arranged to have two large scale perpendicular gradients mixed by the turbulent velocity field. In this case the synthetic pigeon will learn what is the angular profile of odors at

home that will encode for the direction of the large scale odor gradients. The homing phase will then start from a point distant from the home site. A preliminary setup is shown in Fig. 4.11B. The green and red gradient point in two different cardinal points, North and East respectively. Obviously the solution we found for the POMDP in the Gaussian uncorrelated odor signals is optimal in that framework and it only allowed us to qualitatively look at behaviors already highlighted in experiments. We hope that the navigation in the turbulent environment will allow us to make some more quantitative *in silico* experiments to validate the findings of the one dimensional case. We also would like to see how the resulting trajectories correlate with real data. The idea that memory and learning are managed by organisms in a Bayesian fashion is not new. Experiments have shown that animals learn and make decisions based on the strength of their beliefs, the reliability of cues, and the likelihood of outcomes integrating past and present information. [21**?** –23]. Olfactory navigation in pigeons still constitutes an open problem with fascinating implications. In a simplified situation our approach connected memory and learning directly to optimal decision-making. We hope that the algorithmic approach we propose in this work could open a new direction to investigate the phenomenon.

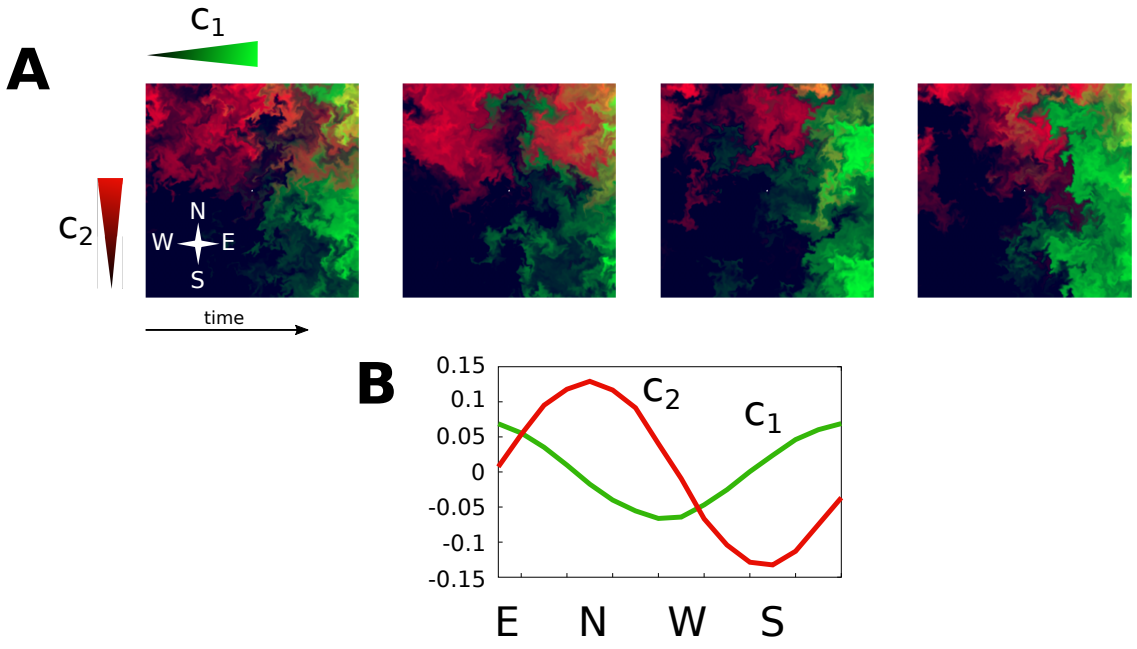


Figure 4.11: **Turbulent odor landscape A)** Four snapshots of a turbulent environment with two odor gradients. The green odor presents a W-E positive gradient perpendicular to the positive S-N one of the red odor. The home site is in the center. **B)** Concentration profiles for the two odors. The reference concentration is $\bar{c} = 0$ for both the odors. The red concentration has a peak in the direction North and the green one peaks at East.

Appendix

B.1 One-dimensional case

In this section we show the detailed solution of the Bellman equation in the one - dimensional case with known and unknown gradient. As mentioned in the main text the goal is to locate the zero of the function defined by $f(x) = gx \equiv s$ with Gaussian distributed random variables representing the odor.

B.1.1 The case of fixed gradient: solution of the Bellman equation

In the case of fixed gradient the state is represented only by s.

The POMDP is defined by the model

$$p(s'|s,a) = \delta(s' - s - ga). \tag{B.1}$$

The likelihood is

$$\ell(c|s) = N(s, \sigma_c^2) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\Big[- \frac{(c-s)^2}{2\sigma_c^2}\Big], \tag{B.2}$$

and the belief of being in state s takes the form

$$b(s) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[- \frac{(s-\mu)^2}{2\sigma^2}\Big]. \tag{B.3}$$

We also fix the actions to be $a = \pm\delta$.

The last ingredient to define is the reward function

$$r(s) = -s^2. \tag{B.4}$$

The belief update gives a new Gaussian $N(\mu', \sigma'^2)$

$$b'(s'|a,c) = \frac{\int ds b(s) \ell(c|s') p(s'|s,a)}{\ell(c|a)} = N(\mu', \sigma'^2) \tag{B.5}$$

with

$$\begin{aligned} \mu' &= \sigma'^2 \Big[\frac{(\mu + ga)}{\sigma^2} + \frac{c}{\sigma_c^2} \Big] \\ \frac{1}{\sigma'^2} &= \frac{1}{\sigma^2} + \frac{1}{\sigma_c^2}. \end{aligned} \tag{B.6}$$

If we define $\sigma_c^2/\sigma^2 \equiv T$ we can rewrite the update equations for μ and σ^2 as

$$\begin{aligned}\mu' &= \frac{T}{T+1}(\mu + ga) + \frac{c}{T+1} = \mu + ga + \frac{c - (\mu + ga)}{T+1} \\ T' &= T + 1.\end{aligned} \tag{B.7}$$

The *evidence* is

$$\ell(c|a) = \int dsds' b(s)\ell(c|s')p(s'|s,a) = N\Big(\mu + ga, \frac{T+1}{T}\sigma_c^2\Big). \tag{B.8}$$

The reward averaged over the belief is

$$\overline{r} = \int dsr(s)b(s) = -\mu^2 - \frac{\sigma_c^2}{T} \tag{B.9}$$

We can now write the optimal Bellman equation

$$V(\mu, T) = -\mu^2 - \sigma_y^2 T^{-2} + \gamma \max_a \Big[\int dc\ell(c|a)V(\mu', T')\Big] \tag{B.10}$$

If we make the ansatz

$$V(\mu, T) = -b\mu^2 - d\mu - \sigma_c^2\phi(T) \tag{B.11}$$

and substitute it inside the equation above we have

$$\begin{aligned}-b\mu^2 - d\mu - \sigma_c^2\phi(T) = -\mu^2 - \frac{\sigma_c^2}{T} \\ -\gamma \min_{a=\pm\delta}\Big[b\int dc\,\ell(c|a)\,\mu'^2 + \\ + d\int dc\,\ell(c|a)\,\mu'\Big] + \\ + \sigma_c^2\phi(T').\end{aligned}$$

Using (B.7) we get

$$\begin{aligned}-b\mu^2 - c\mu - \sigma_c^2\phi(T) = &-\mu^2 - \frac{\sigma_c^2}{T} - \gamma \min_{a=\pm\delta}\Big[b(\mu + ga)^2 + d(\mu + ga)\Big] + \\ &- \sigma_c^2\phi(T') - b\frac{\sigma_c^2}{T(T+1)}.\end{aligned}$$

Equating the terms of order μ^2 gives

$$b = 1/(1-\gamma)$$

In the min operation $a = +\delta$ is optimal if

$$b(\mu+g\delta)^2+d(\mu+g\delta)<b(\mu-g\delta)^2+d(\mu-g\delta)$$
$$\downarrow$$
$$\mu<-\frac{d}{2b}.$$

with $d=\frac{\gamma}{(1-\gamma)^2}g\delta$ obtained equating linear terms in μ. Finally, the optimal policy we obtain is

$$a^*=\begin{cases}+\delta, & \mu<-\frac{\gamma g\delta}{2}\\ -\delta, & \mu>\frac{\gamma g\delta}{2}\end{cases}$$

with μ that is updated by using new observation as

$$\begin{aligned}\mu' &= \mu+ga+\alpha'_T\big[c-(\mu+ga)\big]\\ \alpha'_T &= \frac{1}{T+1}.\end{aligned}\tag{B.12}$$

In this way the choice of the action is not completely reactive according to the new observation but it integrates the odor signal to have a more precise estimate of the true odor concentration along the trajectory thus giving a more precise hint on the position with respect to the target.

B.1.2 The case of unknown gradient

In this section we show the details of the solution of the Bellman equation

$$V(\overline{\mu},M)=\overline{r}+\gamma\max_a\left[\int dc\,\ell(c|a)\,V(\overline{\mu}',M')\right]\tag{B.13}$$

to obtain eq. (12) of the main text.

The update rule for the covariance matrix M of the belief are

$$\begin{aligned}M'^{ss} &= M^{ss}+\sigma_y^{-2}\\ M'^{sg} &= M^{sg}-aM^{ss}\\ M'^{gg} &= M^{gg}+a^2M^{ss}-2aM^{sg}\end{aligned}\tag{B.14}$$

and can be easily obtained by comparing the terms in s^2,sg and g^2 in eq. (4).

The linear terms in s, g give respectively the two equations

$$\begin{aligned}M'_{ss}\mu'_s+M'_{sg}\mu'_g &= M_{ss}\mu_s+M_{sg}\mu_g+\frac{c}{\sigma_c^2}\\ M'_{gs}\mu'_s+M'_{gg}\mu'_g &= M_{gs}\mu_s+M_{gg}\mu_g-aM_{ss}\mu_s-aM_{sg}\mu_g\end{aligned}$$

Using the update relations for M' we can rewrite this two equations as

$$\begin{aligned} M'_{ss}[\mu'_s - (\mu_s + a\mu_g)] + M'_{sg}(\mu'_g - \mu_g) &= \frac{1}{\sigma_c^2}[c - (\mu_s - a\mu_g)] \\ M'_{sg}[\mu'_s - (\mu_s + a\mu_g)] + M'_{gg}(\mu'_g - \mu_g) &= 0 \end{aligned}$$

thus

$$\begin{pmatrix} \mu'_s - (\mu_s + a\mu_g) \\ \mu'_g - \mu_g \end{pmatrix} = M'^{-1} \begin{pmatrix} \frac{c-(\mu_s+a\mu_g)}{\sigma_c^2} \\ 0 \end{pmatrix} \tag{B.15}$$

where

$$M'^{-1} = \begin{pmatrix} M'_{ss} & M'_{sg} \\ M'_{gs} & M'_{gg} \end{pmatrix}^{-1} = \frac{1}{\det M'} \begin{pmatrix} M'_{gg} & -M'_{sg} \\ -M'_{gs} & M'_{ss} \end{pmatrix}. \tag{B.16}$$

Combining the previous two set of equations we obtain exactly the two equation in the main text

$$\begin{aligned} \mu'^s &= \mu^s + a\mu^g + \alpha[c - (\mu^s + a\mu^g)] \\ \mu'^g &= \mu^g - \beta[c - (\mu^s + a\mu^g)] \end{aligned} \tag{B.17}$$

with learning rates

$$\alpha = \frac{M'^{gg}}{\sigma_y^2 \det M'} = \frac{1}{\sigma_y^2}(M'^{-1})^{ss} \ , \quad \beta = -\frac{M'^{sg}}{\sigma_y^2 \det M'} = \frac{1}{\sigma_y^2}(M'^{-1})^{sg}. \tag{B.18}$$

We can now solve the optimal Bellman equation. A relation that will be useful in the following is

$$\det M' = \det M + \frac{M'_{gg}}{\sigma_c^2}. \tag{B.19}$$

to make the notation more readable we define

$$\begin{aligned} z &\equiv \mu_s + a\mu_g \\ \overline{\sigma}^2 &\equiv \sigma_c^2 \frac{\det M'}{\det M} \end{aligned}$$

The *evidence* is

$$\begin{aligned} \ell(c|a) &= \int ds' dg b(s' - ga, g)\ell(c|s') \\ &= \frac{1}{\sqrt{2\pi\overline{\sigma}^2}} \exp\Big[- \frac{(c-z)^2}{2\overline{\sigma}^2} \Big] \end{aligned} \ . \tag{B.20}$$

One term of the Bellman equation presents the $\int dy f(y|a)V^*(\overline{\mu}', M') \equiv \mathbb{E}_{\ell(c|a)}[V'^*]$ that is the average value of the value function with respect to the evidence.

We can now compute $\mathbb{E}_{\ell(c|a)}[\mu_s'^2]$, $\mathbb{E}_{\ell(c|a)}[\mu_g'^2]$ and $\mathbb{E}_{\ell(c|a)}[\mu_s'\mu_g']$.

$$\begin{aligned}
\mathbb{E}_{\ell(c|a)}[\mu_s'^2] &= \int dc\ \mu_s'^2\ \ell(c|a) \\
&= \int dc\big[z^2 + \alpha^2(c-z)^2 + \alpha z(c-z)\big] \\
&= z^2 + \alpha^2\overline{\sigma}^2 \\
&= (\mu_s + a\mu_g)^2 + M'_{gg}\Big(\frac{1}{\det M} - \frac{1}{\det M'}\Big) \\
&= -M'^{-1}_{ss} + \mu_s^2 + M^{-1}_{ss} + a^2[\mu_g^2 + M^{-1}_{gg}] + 2a[\mu_s\mu_g + M^{-1}_{sg}]
\end{aligned} \tag{B.21}$$

from which

$$\mathbb{E}_{\ell(c|a)}[\mu_s'^2 + M'^{-1}_{ss}] = \mu_s^2 + M^{-1}_{ss} + a^2[\mu_g^2 + M^{-1}_{gg}] + 2a[\mu_s\mu_g + M^{-1}_{sg}]. \tag{B.22}$$

The average value of $\mu_g'^2$ over the evidence becomes

$$\begin{aligned}
\mathbb{E}_{\ell(c|a)}[\mu_g'^2] &= \mu_g^2 + \beta^2\overline{\sigma}^2 \\
&= \mu_g^2 + M'_{ss}\Big(\frac{1}{\det M} - \frac{1}{\det M'}\Big) - \frac{1}{\sigma_c^2 \det M}. \\
&= -M'^{-1}_{gg} + \mu_g^2 + M^{-1}_{gg}
\end{aligned} \tag{B.23}$$

from which

$$\mathbb{E}_{\ell(c|a)}[\mu_g'^2 + M'^{-1}_{gg}] = \mu_g^2 + M^{-1}_{gg}. \tag{B.24}$$

For the average value of $\mu_s'\mu_g'$ we can proceed in the same way to obtain

$$\mathbb{E}_{\ell(c|a)}[\mu_s'\mu_g' + M'^{-1}_{sg}] = \mu_s\mu_g + M^{-1}_{sg} + a\Big(\mu_g^2 + M^{-1}_{gg}\Big). \tag{B.25}$$

If we now define

$$V(\overline{\mu}, M) = -A\big[\mu_s^2 + M^{-1}_{ss}\big] - B\big[\mu_s\mu_g + M^{-1}_{sg}\big] - D\big[\mu_g^2 + M^{-1}_{gg}\big]$$

the expression $\mathbb{E}_{\ell(c|a)}[V'^*]$ inside the $\max_a$ in the Bellman equation takes the form

$$\mathbb{E}_{\ell(c|a)}[V'^*] = -A[\mu_s^2 + M^{-1}_{ss}] - (a^2A + aB + D)[\mu_s\mu_g + M^{-1}_{sg}] - (2aA + B)[\mu_g^2 + M^{-1}_{gg}].$$

The `l.h.s.` and `r.h.s.` of the Bellman equation are equal if

$$A = \frac{1}{1-\gamma} \quad B = \frac{2\gamma a_*}{(1-\gamma)^2} \quad D = \frac{a_*^2\gamma(1+\gamma)}{(1-\gamma)^3}$$

Since we are in the one dimensional case and the only coefficient linear in the action is B we have that the $\max_a$ is selected by $\mathbb{E}_b[sg] = \mu_s\mu_g + M_{sg}^{-1}$, thus obtaining equation 13.

B.2 The general case of multiple odors

In this case following the same steps of the previous section we derive the solution for the value function in the multidimensional and multiple odors case. As explained in the main text the goal is to find the zero of the multidimensional function $F(x) = gx$ by means of observations made about n Gaussian distributed and independent odors. $x \in \mathbb{R}^d$, $F \in \mathbb{R}^n$ and g is a $n \times d$ matrix. In this case the action a is a $d-$dimensional vector.

The posterior belief is

$$b'(s', g|c, a) = \frac{b(s' - ga, g)\ell(c|s')}{\ell(c|a)} \tag{B.26}$$

with $b(s, g)$, $\ell(c|s)$ specified by eqs.(18) and (17) of the main text, respectively.

Collecting the quadratic terms in s and g we obtain, as in the one-dimensional case

$$\begin{aligned} M'_{ss} &= M_{ss} + K \\ M'_{sg} &= M_{sg} - M_{ss} \otimes a^T \\ M'_{gg} &= M_{gg} - M_{sg} \otimes a - M_{gs} \otimes a^T + M_{ss} \otimes (aa^T) \end{aligned} \tag{B.27}$$

where $\otimes$ here represents the usual Kronecker product and K^{-1} is the $n \times n$ covariance matrix for the observations.

Collecting the linear terms in s and g results in

$$\begin{aligned} M'_{ss}\mu'_s + M'_{sg}\mu'_g &= M_{ss}\mu_s + M_{sg}\mu_g + Ky \\ M'_{gs}\mu'_s + M'_{gg}\mu'_g &= M_{gs}\mu_s + M_{gg}\mu_g - (M_{ss} \otimes a)\mu_s - (M_{sg} \otimes a)\mu_g \end{aligned}$$

Using the relations for the covariance matrix M' we can rewrite the relations above as

$$\begin{aligned} M'_{ss}[\mu'_s - (\mu_s + \mu_g a)] + M'_{sg}(\mu'_g - \mu_g) &= K[c - (\mu_s - \mu_g a)] \\ M'_{sg}[\mu'_s - (\mu_s + \mu_g a)] + M'_{gg}(\mu'_g - \mu_g) &= 0 \end{aligned}$$

from which the two equations (23) for μ'_s and μ'_g follow.

The rationale we followed to solve the Bellman equation is exactly the same as the one exposed in the previous section.

To solve

$$V = \bar{r} + \gamma \max_a \mathbb{E}_{\ell(c|a)}[V']$$

it is sufficient to note that

$$\int dc\ \ell(c|a)\left[\mu_s'^T\mu'_s + \operatorname{tr} M'^{-1}_{ss}\right] = (\mu_s + \mu_g a)^T(\mu_s + \mu_g a) + \operatorname{tr} M_{ss}^{-1} + 2\operatorname{tr}(M_{sg}^{-1}a) + \operatorname{tr}(a^T M_{gg}^{-1}a)$$
$$\int dc\ \ell(c|a)\left[\mu_s'^T\mu'_g a + \operatorname{tr}(M'^{-1}_{sg}a)\right] = (\mu_s + \mu_g a)^T\mu_g a + \operatorname{tr}(M_{sg}^{-1}a) + \operatorname{tr}(a^T M_{gg}^{-1}a)$$
$$\int dc\ \ell(c|a)\left[a^T\mu_g'^T\mu'_g a + \operatorname{tr}(a^T M'^{-1}_{gg}a)\right] = a^T\mu_g^T\mu_g a + \operatorname{tr}(a^T M_{gg}^{-1}a).$$

For this reason the ansatz

$$V(\overline{\mu}, M) = -A\left[\mu_s^T\mu_s + \operatorname{tr} M_{ss}^{-1}\right] - B\left[\mu_s^T\mu_g a + \operatorname{tr}(M_{sg}^{-1}a)\right] - C\left[a^T\mu_g^T\mu_g a + \operatorname{tr}(a^T M_{gg}^{-1}a)\right]$$

solves the Bellman equation if

$$A = \frac{1}{1-\gamma} \quad B = \frac{2\gamma}{(1-\gamma)^2} \quad D = \frac{\gamma(1+\gamma)}{(1-\gamma)^3}.$$

The optimal policy is then defined by

$$a^* = \arg\max_a V(\bar{\mu}, M)$$

where a belongs to a set of discrete actions.

In this case the term involving the gradients influence the choice of the optimal action at variance from what we found in the one-dimensional situation where that term was contributing just a constant shifting the baseline of the value function.

B.2.1 Recursive relations

In this section we derive in the general case of n odors in d dimension, the recursive relations for the covariance matrix M and consequently for the learning rates α and β.

The two learning rates are defined by

$$\alpha = \frac{1}{\sigma_c^2}(M'^{-1})^{ss}\ , \quad \beta = \frac{1}{\sigma_c^2}(M'^{-1})^{sg}. \tag{B.28}$$

We now find a recursive expression for M'^{-1}.

$$M = \begin{bmatrix} M_{ss} & M_{sg} \\ M_{gs} & M_{gg} \end{bmatrix}^{-1} \equiv \begin{bmatrix} M_{ss}^{-1} & M_{sg}^{-1} \\ M_{gs}^{-1} & M_{gg}^{-1} \end{bmatrix}$$
$$= \begin{bmatrix} (M_{ss})^{-1} + (M_{ss})^{-1} M_{sg} Q^{-1} M_{gs} (M_{ss})^{-1} & -(M_{ss})^{-1} M_{sg} Q^{-1} \\ -Q^{-1} M_{gs} (M_{ss})^{-1} & Q^{-1} \end{bmatrix}$$

where

$$Q = M_{gg} - M_{gs}(M_{ss})^{-1} M_{sg}. \tag{B.29}$$

In this case M is a $(n+1)d \times (n+1)d$ matrix with a $n \times n$ block M_{ss}, $n \times nd$ block M_{sg} and $nd \times nd$ block M_{gg}. $M_{sg} = M_{gs}^T$.

If a $t = 0$ is $M = 0$ we have

$$M_{ss}^t = tK$$
$$M_{sg}^t = -K \otimes \Delta_t$$

with $\Delta_t = \sum_{\tau=1}^{t-1} \tau a_\tau = \Delta_{t-1} + (t-1)a_t$. Using these expressions in the definition of M_{gg}^{t+1} we can rewrite

$$M_{gg}^{t+1} - M_{gg}^t = \frac{1}{t} K \otimes \left[\Delta_{t+1} \Delta_{t+1}^T + \Delta_t \Delta_t^T \right].$$

Performing the telescopic sum we get

$$M_{gg}^t = K \otimes H_{t-1} + \frac{1}{t-1} K \otimes (\Delta_t \Delta_t^T)$$

where H_t is a symmetric $d \times d$ matrix $H_t = \sum_{\tau=2}^{t} \frac{1}{\tau(\tau-1)} \Delta_\tau \Delta_\tau^T = H_{t-1} + \frac{1}{t(t-1)} \Delta_t \Delta_t^T$.

Now that we have the recursive relations for each block of the matrix M we can compute Q that is then necessary to compute the inverse of each block of M. Given the definition of Q and the recursive expressions for M we get

$$Q_t = K \otimes H_t \to Q_t^{-1} = K^{-1} \otimes H_t^{-1}$$

Eventually the learning rates of eq. (23) have the following expressions

$$\alpha_t = M_{ss}^{-1} K = \frac{1}{t} \left(1 + \frac{1}{t} \Delta_t^T H^{-1} \Delta_t \right)$$
$$\beta_t = M_{gs}'^{-1} K = \frac{1}{t} H_t^{-1} \Delta_t.$$

In the case $n = 1$ and $d = 1$ we have $H \to \sigma_c^{-2}$ and the expression for M^{-1} is the usual inverse of a 2×2 matrix.

Bibliography

[1] Shaffer, S. A., Tremblay, Y., Weimerskirch, H., Scott, D., Thompson, D. R., Sagar, P. M., Moller, H., Taylor, G. A., Foley, D. G., Block, B. A., and Costa, D. P. Migratory shearwaters integrate oceanic resources across the pacific ocean in an endless summer. *Proceedings of the National Academy of Sciences*, 103 (34):12799–12802, aug 2006.

[2] Wallraff, H. G. *Avian navigation: Pigeon homing as a paradigm*. 2005.

[3] Mouritsen, H. Long-distance navigation and magnetoreception in migratory animals. *Nature*, 558(7708):50–59, 2018.

[4] Wallraff, H. G. Pigeon homing from unfamiliar areas: An alternative to olfactory navigation is not in sight. *Communicative and Integrative Biology*, 2014.

[5] Keeton, W. T. Orientation by pigeons: Is the sun necessary? *Science*, 165 (3896):922–928, 1969.

[6] Keeton, W. T. Magnets Interfere with Pigeon Homing. *Proceedings of the National Academy of Sciences*, 68(1):102–106, 1971.

[7] Walcott, C. and Green, R. P. Orientation of homing pigeons altered by a change in the direction of an applied magnetic field. *Science*, 184(4133):180–182, 1974.

[8] Ioale, P. and Guidarini, D. Methods for producing disturbances in pigeon homing behaviour by oscillating magnetic fields. *Journal of Experimental Biology*, 120:109–120, 1985.

[9] Papi, F., Fiore, L., Fiaschi, V., and Benvenuti, S. The influence of olfactory nerve section in the homing capacity of carrier pigeons. *Italian Journal of Zoology*, 1971.

[10] Gagliardo, A. Forty years of olfactory navigation in birds. *Journal of Experimental Biology*, 2013.

[11] Patzke, N., Manns, M., Güntürkün, O., Ioalè, P., and Gagliardo, A. Navigation-induced ZENK expression in the olfactory system of pigeons (Columba livia). *European Journal of Neuroscience*, 31(11):2062–2072, 2010.

[12] Papi, F., Ioalé, P., Fiaschi, V., Benvenuti, S., and Baldaccini, N. E. Olfactory navigation of pigeons: The effect of treatment with odorous air currents. *Journal of Comparative Physiology ??? A*, 94(3):187–193, 1974.

[13] Wallraff, H. G. Simulated Navigation Based on Assumed Gradients of Atmospheric Trace Gases (Models on Pigeon Homing, Part 2). *Journal of Theoretical Biology*, 1989.

[14] Wallraff, H. G. Simulated navigation based on observed gradients of atmospheric trace gases (Models on pigeon homing, Part 3). *Journal of Theoretical Biology*, 205(1):133–145, 2000.

[15] Boström, J. E., Åkesson, S., and Alerstam, T. Where on earth can animals use a geomagnetic bi-coordinate map for navigation? *Ecography*, 35(11):1039–1047, 2012.

[16] Wallraff, H. G. and Andreae, M. O. Spatial gradients in ratios of atmospheric trace gases: A study stimulated by experiments on bird navigation. *Tellus, Series B: Chemical and Physical Meteorology*, 52(4):1138–1157, 2000.

[17] Wallraff, H. G. Ratios among atmospheric trace gases together with winds imply exploitable information for bird navigation: A model elucidating experimental results. *Biogeosciences*, 10(11):6929–6943, 2013.

[18] Wallraff, H. G. Simulated navigation based on unreliable sources of information (models on pigeon homing. Part 1). *Journal of Theoretical Biology*, 137(1):1–19, 1989.

[19] Balkovsky, E. and Shraiman, B. I. Olfactory search at high Reynolds number. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12589–93, 2002.

[20] Vergassola, M., Villermaux, E., and Shraiman, B. I. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, 2007.

[21] Biernaskie, J. M., Walker, S. C., and Gegear, R. J. Bumblebees Learn to Forage like Bayesians. *The American Naturalist*, 174(3):413–423, 2009.

[22] Foley, B. R. and Marjoram, P. Sure enough: efficient Bayesian learning and choice. *Animal Cognition*, 20(5):867–880, 2017.

[23] J. Valone, T. Are animals capable of bayesian updating? an empirical review. *Oikos*, 112(2):252–259, 2006.

Chapter 5

Optimality of trace fossils

A look at their shapes through decision making theory

1 Introduction

Most trace fossils represent what remains of the behavior of organisms that are rarely preserved as body fossils. For this reason the morphology of the traces serves as a proxy for the ethology of these organisms at the time the trace was produced. The study of fossil traces left by soft-bodied organisms goes under the name of *ichnology*. Ichnologists highlighted a very interesting point: the behavior exhibited by trace fossils shows a diversification over geological time, showing an evolution towards more optimized search and foraging patterns and the development of new strategies. The first foraging trails present in the records show tracks that often cross themselves, and indicate relatively crude foraging strategies (see Fig. 5.1). However, during the "Cambrian explosion" (around 600 millions year ago), more regular foraging patterns were appearing showing spiral fossils or "meandering" trails exhibiting high complexity, compactness and high degree of self-avoidance (see Chapter 3 in [1]). Seilacher [1] suggested that the evolutionary changes in trace fossils, particularly those from the deep sea, involved optimization of feeding behaviour, increase in complexity and compactness.

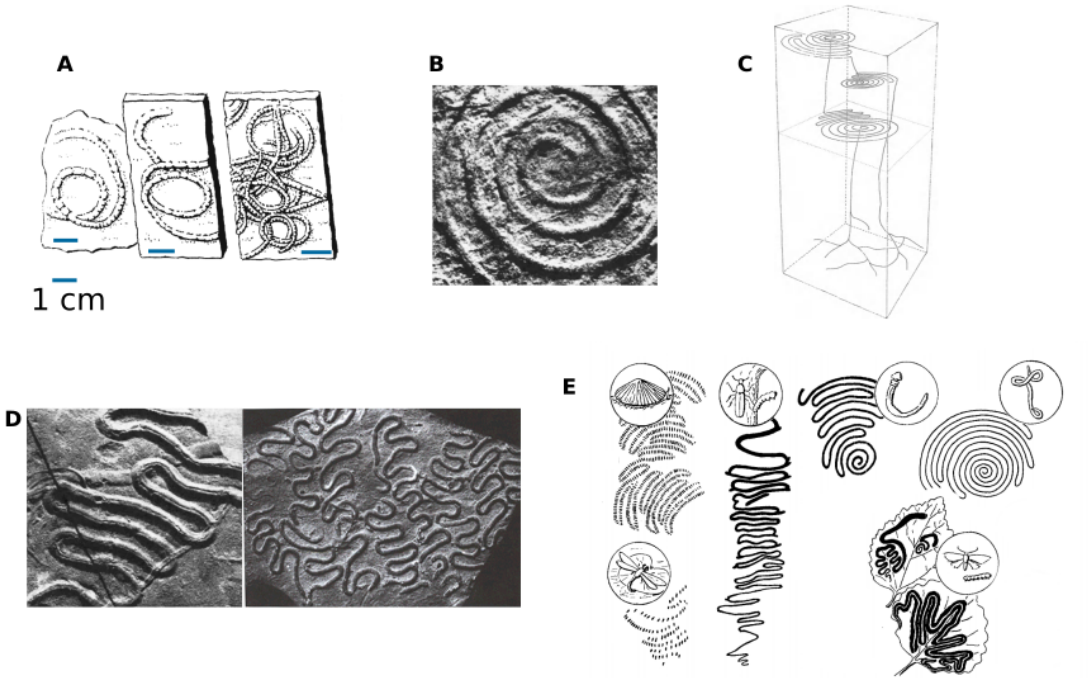


Figure 5.1: **Evolution of strategies.** **A** Early traces (sometimes called "scribbles") showing high crossing rates **B** an example of a spiral **C** The planar nature of the spiraling trajectories is part of network of spirals connected by straight trajectories **D** Examples of meanders **E** *Modern* meanders and spirals trajectories produced by various species such as limpets (grazing algal films), dragonflies and bark beetles (laying their eggs), acorn worms and Praonis (in the sea bed), moths (*Ogmograptis* lying eggs on leaf or barks). [adapted from [2, 3]]

The analysis of the fossil traces revealed that the behavior encoded in the trajectories could be reproduced by using three simple rules (see Fig. 5.2)

1. *Phobotaxis* that forbids the worm to cross its own trail (or any other trail)
2. *Thigmotaxis* used to stay close to an existing trail
3. *Strophotaxis* making the worm reverse direction (U-turn)

In the classification given by Seilacher [3] apart from all the details the categories can be summarized into a few classes. Crawling traces which represent animals moving without necessarily feeding; dwelling traces, interpreted as semi-permanently occupied structures and finally grazing traces generally composed of meanders or spirals and related to animals actively exploiting food resources.

Previous models [4, 5] focused on implementing a hard-coded strategy with the three simple rules (see Fig. 5.2) in an `if-then-else` series of instructions and the

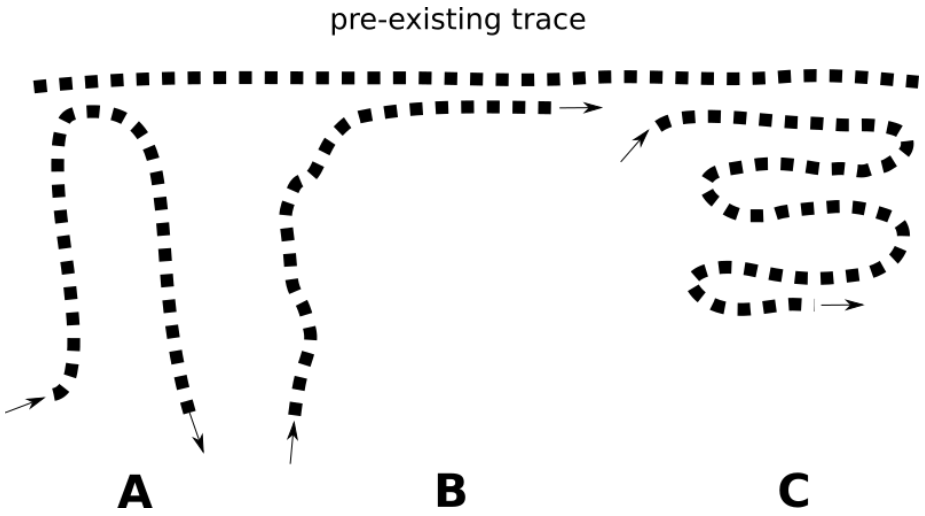


Figure 5.2: **Three basic rules. A** *Phobotaxis* causes the worm to avoid crossing existing tracks **B** *thigmotaxis* forces the worm to stay close to a pre-existing track **C** *strophotaxis* causes the worm to make U-turns

fine tuning of some parameters allowed these models to reproduce meanders and spirals. Hyphotesis about the factors that fostered the development of new and more efficient strategies take into account both the evolution of sensing capabilities of the organisms [6] and the distribution of food in the environment [7]. In another approach Sims et al. (2014) following the optimality argument related to Lévy-like search strategies, show that meanders can be considered optimal in that sense.

Given these premises, the question becomes how the organisms created the patterns, and what kind of algorithms they employed to guide their movement. In the following we will focus on feeding strategies showing how their structure can emerge from optimality principles. We will also focus on how the interaction between sensory capabilities and the surrounding environment shapes the behavior.

2 Scavenging *in silico*

Given the context exposed above our goal was to understand if the rules highlighted as fundamental building blocks in behavior of simple organisms could arise spontaneously without fine tuning of parameters. In particular we were interested in understanding if spirals and meanders, the most common behaviors observed in the fossil traces, could be described as the result of an optimal foraging process.

Following the evidences that early worms had very limited sensory capabilities we wanted to understand how the strategies described above could emerge from both very simple sensing modalities and learning algorithms. Reinforcement learning serves as a natural framework in which to include the problem. Imagine, in fact to

have a given patch of food and an agent that can sense the environment only in the close neighborhood of its body. What kind of strategies will emerge if we ask the agent to learn to act optimally with the request that in a given amount of time it has to eat the most out of the patch ? Following Fig. 5.3 let us suppose that the agent is represented as a disk of radius R. Moreover its sensitivity goes up to a distance R_s. We wanted to study two situations in which the agent is only sensitive to directional stimuli (sensing the average direction of food around it Fig. 5.3 A) and to the intensity of the food (Fig. 5.3 B). In both settings the agent represents the environment into a discrete number N of states represented by one of the colored sectors in Fig. 5.3 A. Actions are fixed to be in the set $\mathcal{A} = \{\texttt{go left}, \texttt{go straight}, \texttt{go right}\}$ characterized by a turning angle θ.

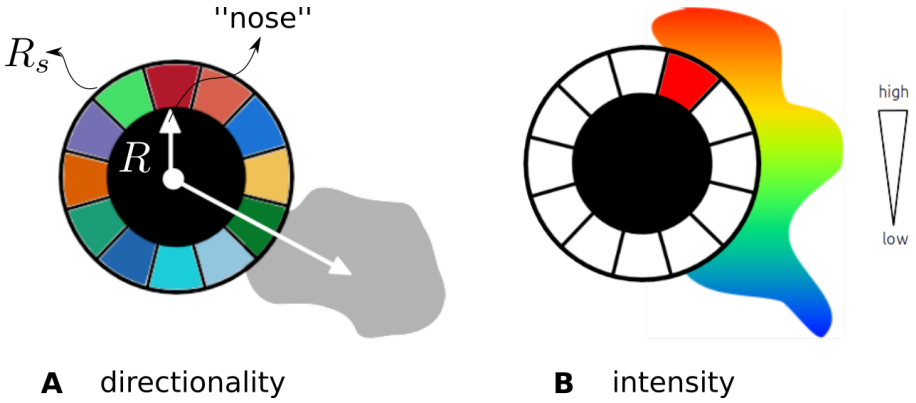


Figure 5.3: **Direction and intensity sensing agents. A)** The agent has size R and sensitivity radius R_s. It represents the environment just looking at the average direction of food in the vicinity of its body **B)** An agent with size and sensitivity radius as in **A** representing the environment according to the intensity of the signal. In this case it is in a gradient (high concentration (red) - low concentration (blue)) and the "sensor" that is activated is the one colored in red.

We give the agent the goal of maximizing the food intake

$$R_T = \sum_{t=0}^{T} \gamma^t r_t \tag{1}$$

in a given amount of time T where r_t is the food eaten at each time step t and γ is the discounting factor. In particular, the simplest situation amounts to consider a greedy agent that wants to maximize the immediate return r_0 corresponding to the case $\gamma = 0$. At the end of each episode of duration T we reset the position of the agent to a random position near the patch with a random orientation of the "nose". Among the different Reinforcement Learning algorithms we use SARSA that has the characteristic of being an on-policy algorithm (see Sec. 2.4). Given the discrete

nature of states and actions the agent will keep a tabular Q-function Q_{sa} representing the value of the state s given the action a. Concerning the policy $\pi(a|s)$ we use what is called an ϵ-greedy strategy where

$$\pi(a|s) = \begin{cases} a^* = \text{argmax}_a\, Q_{sa}, & \text{with probability } 1-\epsilon \\ a \in \mathcal{A} - a^*, & \text{with probability } \epsilon \end{cases} \tag{2}$$

where *epsilon* represents a small probability. This allows the agent to follow the optimal action keeping the exploration of new ones.

3 Results

Spirals and meanders

The first result we discuss is related to the agent that directionally senses the food around. We tested the algorithm in two situations in which the agent has to exploit a circular and a squared patch of homogeneous distribution of food both with linear size L. We fix the the radius of sensitivity to be $R_s = 1.3R$ with the size of the agent $R = \frac{1}{20}L$. The number angular states is $N = 6$ providing the agent a very raw representation of the surrounding food distribution. The learning parameter is fixed at a value $\alpha \sim \frac{1}{T}$ and $\epsilon \sim 10^{-3}$. The episode duration is in both cases $T \sim 1500$ steps. The turning angle is fixed to be $\theta = 20°$. The Q-function is initialized at $Q_{sa} = 0\ \forall s, a$ meaning that the agent is completely naive at the beginning of the training.

The spirals appears in the circular patch at very early learning stages. On average it takes 60 episodes for the agent to learn how to cover the patch using spirals. The spirals that emerge present very high degree of compactness. A more tricky situation is when the agent has to cover a squared patch. In particular the corners represent a difficult configuration where the agent can be stuck (see Fig. 5.5). This is also the reason why the average return is more noisy than in the circular patches. Eventually spirals emerges with the optimal distance ($d \sim 2R$) between successive tracks.

Commenting on these findings we can say that these homogeneous distribution of food and an agent capable of directionally sensing it in the immediate vicinity of its body spontaneously show the emergence of *thigmotaxis*. In the following we will see how the other two *rules*, i.e. *strophotaxis* and *phobotaxis* emerge in the same setting just changing the way the agent represents the environment and introducing gradients in the distribution of food. In this case the agent is sensitive to the intensity of food around its body (see Fig. 5.3 B) instead of representing the direction of food. The other parameters are chosen as in the previous case. Clearly both *strophotaxis* and *phobotaxis* emerges naturally in this context. At the early stages the agent presents close to random strategies (see G_1 to G_3 stages in Fig. 5.6). At stage G_3 the agent starts to understand the structure of the gradient exploiting it upward. In

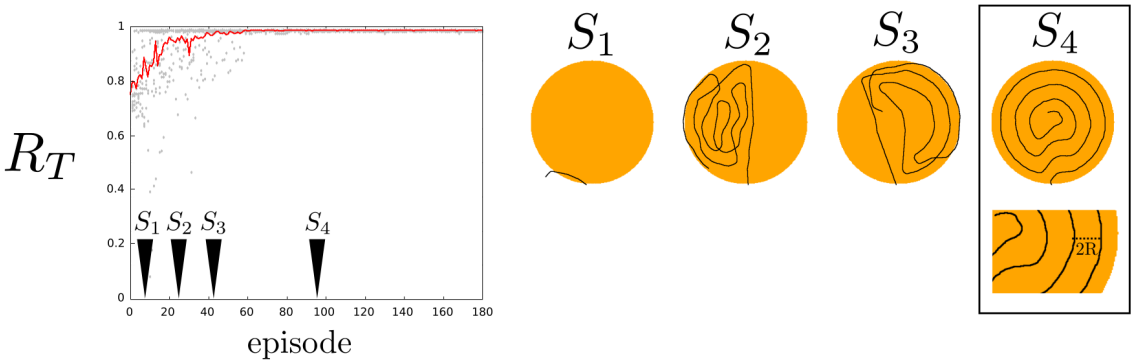


Figure 5.4: **Learning to optimally forage a circular patch.** The plot on the left shows the evolution of the total return as a function of the episode. Grey dots encode for different realizations with their average shown as a red curve. On the right four different stages during the learning. S_2 show a case in which the agent is trapped on one half due to wrong initial actions. Eventually the agent converges to the spiral. The small picture below stage S_4 shows the close packing of the trajectory.

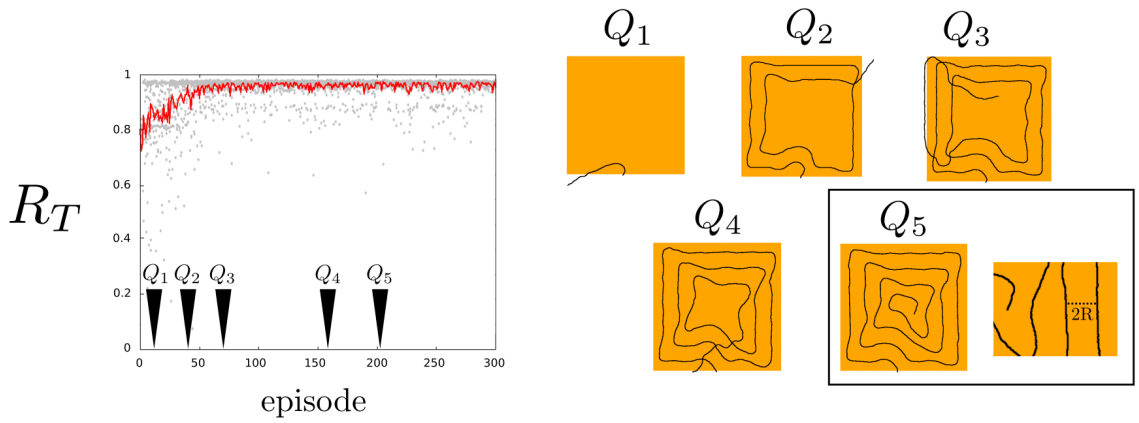


Figure 5.5: **Learning to optimally forage a squared patch.** The plot on the left shows the evolution of the total return as a function of the episode. Grey dots encode for different realizations with their average shown as a red curve. On the right five different stages during the learning. The square patch is more difficult to learn since sometimes the agent can be stuck in the corners then taking the wrong direction ($Q_2\, Q_3\, Q_4$). Eventually the agent converges to the spiral. The small picture near stage Q_5 shows the close packing of the trajectory

the three examples of the final stage G_4 we see the other two fundamental behaviors namely *strophotaxis* and *phobotaxis* as classified in Fig. 5.2.

At this point we ended up with two types of "sensory systems" producing the three basic behaviors in two different environments. *Thigmotaxis* spontaneously

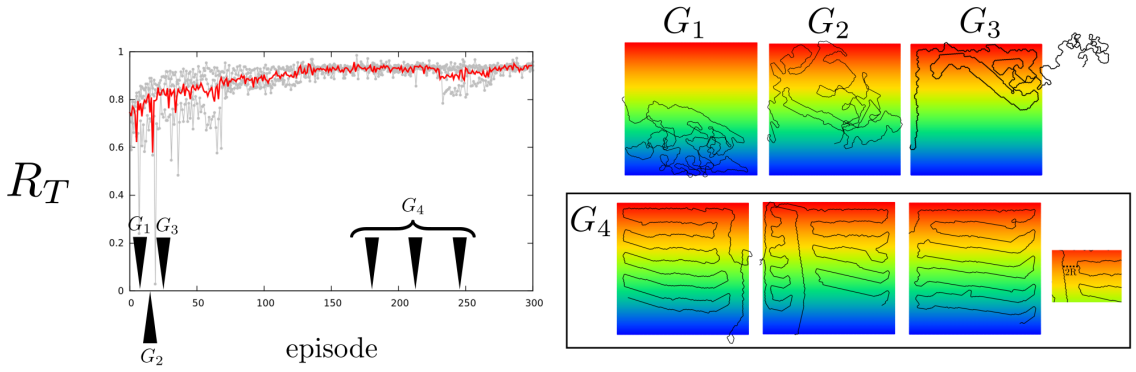


Figure 5.6: **Learning to optimally forage in a gradient.** The plot on the right shows the trend of the total return across episodes of duration $T \sim 1500$ steps. In the early stages (G_1 G_2 G_3) the strategy is close to random. At stage G_3 there is evidence that the agent climbs the gradient. All the three snapshots of stage G_4 highlight the emergence of both *strophotaxis* and *phobotaxis*

emerged in an homogeneous environment in the case of an agent sensitive to the average direction of food. *Phobotaxis* and *strophotaxis* behaviors were generated in the case of a food gradient explored by an agent sensitive to the intensity of food. It was natural to investigate what kind of strategies could emerge in the case of homogeneous environment and an agent representing the food distribution according to its intensity. Fig. 5.7 shows that spirals emerge again as the optimal strategy to exploit the food patch.

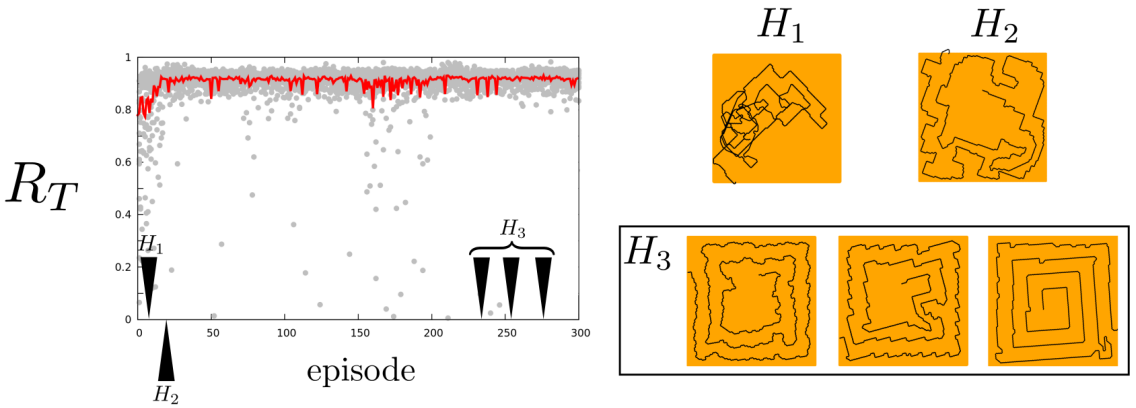


Figure 5.7: **Spirals emerge again.** Left panel shows the average return versus the episode of an agent sensitive to the intensity of surrounding food (see Fig. 5.3). The red line is an average over multiple realizations of the learning process. The arrows on the x-axis pinpoint different episodes with trajectories reported on the right panel. H_1 and H_2 shows early stages trajectories. H_3 refers to three different trajectories found in late episodes.

Localizing a source

The approach we presented can be extended to train an agent to localize a source. In the following, we will take a simple situation in which an agent is able to have a representation of distant food patches. Imagine for example that the agent can track a point that is representative of the food distribution. From very far away we can take this point to be the center of mass of the food distribution giving an indication of the overall position of the region where to point to reach the food. Closer to the food region the agent will be more sensitive to closer patches and the center of mass will turn to be an indicator of the position of these patches (see Fig. 5.8).

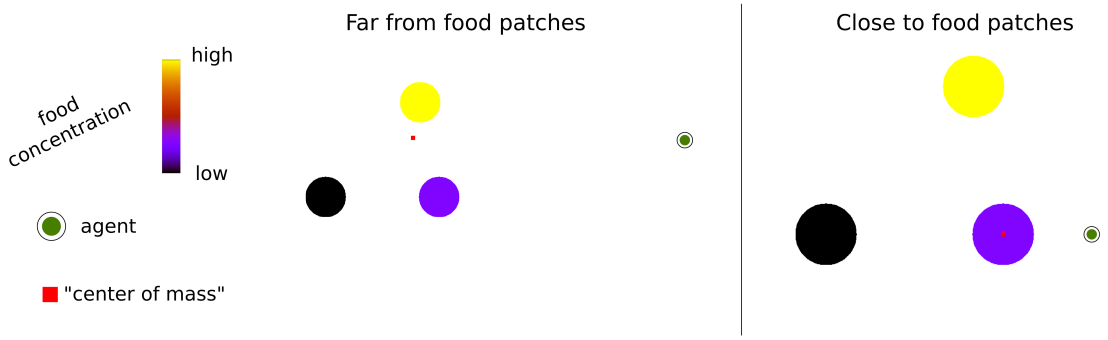


Figure 5.8: **Center of mass representation.** When the agent is far from the region where the food is distributed the center of mass of the distribution is a good indicator. The red point is the center of mass of the three patches closer to the higher concentrated patches (the purple and yellow ones). On approaching the "center of mass" shifts to patches that are closer to the agent. In this case the red point is inside the purple region.

First of all we can train the agent to localize a single patch in a given position. The target is represented as a circle of the same size R of the agent. The episode terminates when the target enters in the sensitivity radius R_s of the agent. The states are defined as the angular position with respect to the agent of the point representing the food distribution. In this case we will take $\gamma = 1$ and the reward will be $r = -1$ if the agent does not reach the target and $r = 100$ if it localizes it. The actions are defined as before to be in the set $\mathcal{A} = \{\texttt{go left}, \texttt{go straight}, \texttt{go right}\}$ with a given angle $\theta = 20°$. The training is summarized in Fig. 5.9. The fastest trajectory is of course the straight line of length L joining the target and the starting point. We track the performance comparing the agent's trajectory (of length d) to L. In the first stages (T_1 in Fig. 5.9) the agent crawls around without intercepting the target. Eventually it finds the target getting high reward. This last event is propagated back to previous states and finally the agent learns to reach the target at each trial (T_2 in Fig. 5.9).

As we said before in the case of multiple targets we can imagine that if the agent

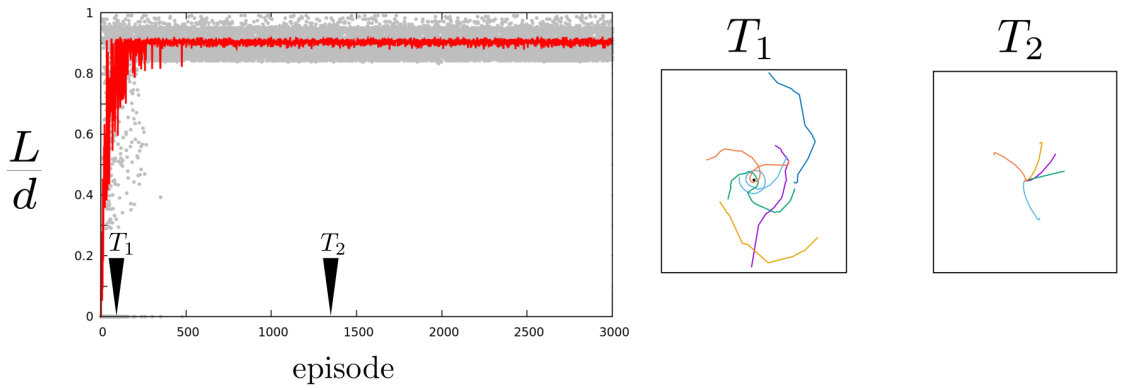


Figure 5.9: **Learning to localize the target.** The plot on the left shows the performance. L is the length of the straight trajectory. d is the length of the actual trajectory. On the right panel we show two stages of the learning. An early phase T_1 where the agent is not able to localize the target. T_2 instead refers to the stage in which the agent has learnt to associate the right sequence of actions to the states. A target of size R equal to the size of the agent is in the middle of the picture. Each trajectory refers to a trial.

is far away from the food distribution the target position can be represented by the center of mass of the food distribution. Instead, when closer to one of the patches the target has to move closer to that particular patch mimicking the fact that the agent will give more credit to closer patches. For this purpose we can think of describing the position of the target $\mathbf{x}_T$ as

$$\mathbf{x}_T = \frac{\sum_i w_i(d_i)\mathbf{x}_i}{\sum_i w_i(d_i)} \tag{3}$$

where the weights $w_i(d_i)$ express how strongly each patch contributes to the center of mass when the agent is at distance d_i from the patches[1]. We can take the weights to be of the form

$$w(d) = f\left[1 + \frac{I}{1 + \exp\left[\lambda(d - d_0)\right]}\right] \tag{4}$$

where f is the intensity of the signal from each patch as perceived from an infinite distance, I is an amplification factor, d_0 is a cutoff distance below which the agent starts to be very sensitive to that patch and λ a decay constant quantifying how fast the "signal" fades away with the distance. The results are presented in Fig. 5.10. In this case we only show the trajectories of the trained agent that follows the policies learnt separately.

[1]We can think of dividing each patch into small parts of intensity f_i. Each part will contribute with a weight $w_i(d_i)$ where d_i is the distance between the agent and the small part.

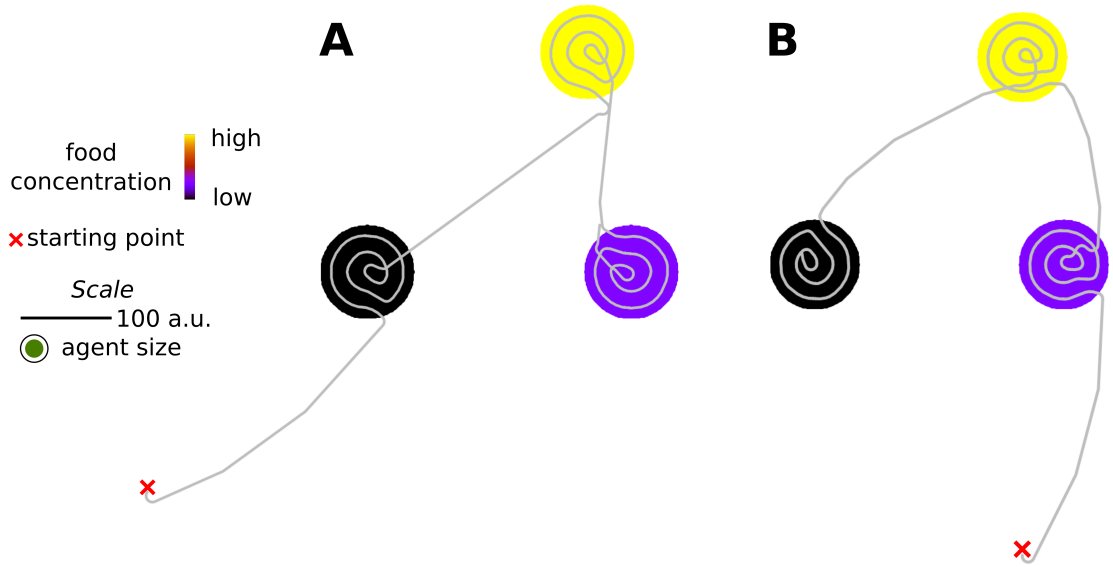


Figure 5.10: **Foraging multiple patches.** Results obtained in the case $\lambda = 0.5$ $d_0 = 200$ and $I = 10^3$. The agent has size $R = 10$ and sensitivity radius $R_s = 1.3R$. When it is not in direct contact with the food it follows the policy to search the target. As soon as it gets to one target it switches to the feeding behavior.

4 Discussion

Raup and Seilacher (1969) were the first to propose a series of hard-coded rules to reproduce trajectories observed in trace fossils. In particular the worm they simulated could move straight ahead, turn toward or away from a preexisting track, or make a full U-turn (they forced a turn of 180° according to some random event). Building on this work Prescott [5] showed that a robot equipped with two lateral sensors gave rise to patterns such as spirals and meanders without specifying the U-turn. In this case a series of predefined events triggered for example a U-turn according to user-defined time-out between one turn and the next. A U-turn was completed as soon as one of the two sensors came in contact with an already deposited track. Moreover to produce spirals and meanders two different architectures were designed (see Fig.3 and Fig.6 in [5]). These approaches were successful attempt to validate the assumption that three very simple rules such as *thigmotaxis*, *strophotaxis* and *phobotaxis* are sufficient to classify and explain the behavioral traits of fossils, drawing possible connections with the evolution of a very simple central nervous system coordinating action selection [6]. Another line of investigation [7] relates the behavior to the characteristics of food distribution. In particular the statement is that the variability in patterns observed in the fossils may represent and be caused by spatial differences in the distribution of food resources.

The algorithmic approach we proposed shows how the strategies resulting in the spirals and meanders often observed in trace fossils can emerge just by requesting the agent to be optimal in the sense of food intake maximization within a given distribution of food. The interaction between sensing capabilities and the surrounding environment is also crucial. As the primitive organisms had limited sensory capabilities, the action selection algorithms must have been fairly simple. In this respect the choice of a Reinforcement Learning approach based on simple learning algorithms able to be implemented in very simple organisms[2] showed to be successful in explaining the observed behaviors. Moreover, we also showed that learning to search a target can be approached in the same framework as well. One aspect that remains to be investigated is the effect of the discounting factor γ (here we considered the case $\gamma = 0$). We found that a greedy agent that wants to maximize just the immediate reward can learn to make spirals and meanders. It can be that the time horizon over which the agent wants to optimize its behavior could play a role in shaping the strategies. Another interesting question could be to understand what kind of search processes appear when an agent faces a more complicated distribution of food with respect to the one it exploited in our examples. Moreover, our approach considered directional sensing in a discretized way. Eucaryotic organisms able to measure spatial gradients and navigate the environment using chemotaxis respond to directional stimuli showing a very well-defined shape of the response function [10]. Another direction we will investigate is to take into account this experimental fact by means of what is called value function approximation. In this kind of reinforcement learning approach the value function is approximated using continuous functions and it would be really interesting to connect the ingredients of this approach to the internal chemical pathways of an organism dedicated to sensing and locomotion.

[2]Here we refer to organisms in general including also the ones that are not equipped with neuron-like structures (see Ref. [9])

Bibliography

[1] Thornhill, R., Nitecki, M., and Kitchell, J. *Evolution of animal behavior: Palentological and field approaches*. 1986.

[2] Bromley, R. *Trace Fossils: Biology, Taxonomy and Applications*. Taylor & Francis, 2012.

[3] Seilacher, A. *Trace Fossil Analysis*. 2007.

[4] Raup, D. and Seilacher, A. Fossil Foraging Behavior : computer simulation. *Science*, 166, 1969.

[5] Prescott T. J. and Ibbotson C. A Robot Trace Maker : Modeling the Fossil Evidence of Early Invertebrate Behavior. *Artificial life*, 3:289–306, 1997.

[6] Prescott, T. J. Forced moves or good tricks in design space? landmarks in the evolution of neural mechanisms for action selection. *Adaptive Behavior*, 15(1): 9–31, 2007.

[7] Plotnick, R. E., Dornbos, S. Q., and Chen, J. Information landscapes and sensory ecology of the Cambrian Radiation. *Paleobiology*, 36(2):303–317, 2010.

[8] Sims, D. W., Reynolds, A. M., Humphries, N. E., Southall, E. J., Wearmouth, V. J., Metcalfe, B., and Twitchett, R. J. Hierarchical random walks in trace fossils and the origin of optimal search behavior. *Proceedings of the National Academy of Sciences*, 111(30):11073–11078, 2014.

[9] Reid, C. R., Garnier, S., Beekman, M., and Latty, T. Information integration and multiattribute decision making in non-neuronal organisms. *Animal Behaviour*, 100:44–50, 2015.

[10] Janetopoulos, C., Ma, L., Devreotes, P. N., and Iglesias, P. A. Chemoattractant-induced phosphatidylinositol 3, 4, 5-trisphosphate accumulation is spatially amplified and adapts, independent of the actin cytoskeleton. *Proceedings of the National Academy of Sciences*, 101(24):8951–8956, 2004.