Constraints and rewards in behavior and optimal decision making

Rethinking optimality in the presence of constraints and the absence of external rewards

Jorge Eduardo Ramírez Ruiz

TESI DOCTORAL UPF / Year 2023

THESIS SUPERVISOR
Dr. Rubén Moreno Bote
Department Tecnologies de la Informació i les Comunicacions

upf. Universitat Pompeu Fabra
Barcelona

A Gael, que siempre encuentres caminos

Acknowledgments

The last four years have been quite a ride for so many reasons, and it is time to thank the people and situations that enriched my life with love, paths, stories and advice. First and foremost, I am grateful for the continued advice from and collaboration with Rubén. I am particularly grateful for the opportunities to pursue my own questions and ideas, as well as the necessary pressure to bring them to fruition (in time!). It has been a great pleasure to share and discuss science, philosophy and life. I am also thankful to the TCN lab, to those who were there when I arrived, Sofía, Devin, Farhad, Alice, Travis, Gabriela, Belén, Ignasi and Alex; as well as to the ones that continue after I am gone, Chiara, Francesco, Fatma, Demetrio and Dmytro. The CBC has been a great and interesting place to develop ideas, I am thankful to the people running things smoothly and to Ege, Gonzalo, Indre, Alice for the coffees, beers and great company.

A big thanks are due to the whole BARCCSYN community, for maintaining a network of like-minded people willing to share science and fun moments. I also highly appreciated the collaboration with Greg deAngelis, Jan Drugowitsch and Becket Ebitz, who have shared a little bit of their time to give me advice and for showing me different ways to do science.

En una nota más personal, resulta increíblemente importante para mí la presencia y amor continuos de mi familia. Gracias a mi mamá y hermano, mi papá y hermanas, y familia extendida, quienes continuamente me muestran su amor y apoyo. Ha sido difícil seguir haciendo esto lejos de ustedes (y más con una pandemia de por medio), pero estoy feliz y agradecido de seguir contando con ustedes a la distancia. Je suis aussi énormement reconnaissant à Rachel, de son amour infini, de son vaste courage, de ses encouragements inconditionnels, de toujours m'aider à simplifier, d'avoir été dans des remerciements déjà trois fois et à chaque fois tu deviens encore plus importante. Y, finalmente, estoy enteramente agradecido por Gael. El doctorado ha encontrado nuevos caminos gracias a ti. Es una alegría indescriptible vivir el amor que intercambiamos, y que ha cambiado nuestras vidas por completo al mismo tiempo que, paradójicamente, nada ha cambiado.

Abstract

The idea of optimal decision-making presupposes certain features about the agent and their environment. This thesis examines two common assumptions in disciplines that study natural and artificial behavior: perfect rationality and reward maximization. Defining or inferring a reward function to maximize can be problematic, especially when one considers the constraints faced by the agents. First, we explore the breadth-depth dilemma, a tradeoff that contrasts superficial versus deep sampling of options by having finite resources. In the models, two major regimes of optimal sample allocation arise as a function of sampling capacity, offering alternative ways to understand "suboptimal" behavior. Additionally, we propose a novel intrinsic motivation approach based on occupying as many paths in the environment as possible, using rewards as means rather than the goal. Agents can thus attach meaning to reward, and develop diverse yet goal-directed behaviors. This approach presents novel opportunities to understand fluid, naturalistic behavior.

Keywords: decision making, optimality, constrained optimization, bounded rationality, breadth–depth tradeoff, intrinsic motivation, reward hypothesis, entropy, reinforcement learning, goal-directed behavior

Resumen

La noción de decisiones óptimas presupone algunas características del agente y su entorno. Esta tesis examina dos suposiciones comunes en diversas disciplinas que estudian comportamiento natural y artificial: racionalidad perfecta y maximización de recompensas. Definir o inferir una función de recompensa a maximizar puede ser problemático, especialmente cuando uno considera las constricciones a las que el agente se enfrenta. Primeramente, esta tesis explora el dilema amplitud–profundidad, un balance que contrasta un muestreo superficial contra uno profundo de las opciones a elegir al tener recursos limitados. En nuestros modelos, dos regímenes principales emergen para la distribución óptima de recursos en función de la capacidad de muestreo, lo cual ofrece alternativas para entender algunos comportamientos "subóptimos". Adicionalmente, se propone una perspectiva de motivación intrínseca basada en la ocupación máxima de trayectorias en el entorno, usando las recompensas como medio y no como fin. Los agentes pueden así asignar un significado a las recompensas, y desarrollan comportamientos variables al mismo tiempo que orientado a metas. Este enfoque ofrece nuevas oportunidades de entender comportamientos naturales y fluidos.

Palabras clave: toma de decisiones, optimalidad, optimización con restricciones, racionalidad limitada, balance amplitud–profundidad, motivación intrínseca, hipótesis de recompensa, entropía, aprendizaje por refuerzos, comportamiento orientado a metas

Preface

During these last four years, I have been exposed to a torrent of diverse ideas, all pertaining to some degree to the study of intelligence and behavior. Although I participated in conferences, courses and summer schools that were technically about *neuroscience*, its multidisciplinary nature helped me interact with all sorts of people and backgrounds ranging from artificial intelligence and computer science, through cognitive and computational neuroscience, all the way to computational biology, ethology and philosophy. Throughout all of these interactions, specifically for the field of decision making, I never stopped noticing the big question of the agential homunculus (as compared to the cortical homunculus); the idea that there is always an agent, either distributed in a circuit or localized abstractly, that makes decisions. It is always either side-stepped, ignored or acknowledged to be a hard question. To me, this big unknown has driven my continued interest in the field of decision making, helping me elucidate two of my big questions: What does the concept of agency get us? And how can we understand the idea of having the ability to choose?

The notions of decision making and behavior are intricately linked by the conception of agency, *i.e.* the ability to produce actions. While behavior is usually thought simply as something that organisms *do*, and decision making usually implies a deliberation process, there are numerous common perspectives that one can take in order to study both, and I have learned the value of taking a pluralistic view in science. Therefore, in the Introduction, I review two pluralistic approaches to the study of biological phenomena: Tinbergen's four questions and Marr's levels of analysis. I then present the field of Bounded Rationality through the lens of these frameworks, given its major relevance in the works presented in the thesis.

One core concept that I have encountered several times in my trajectory is the notion of *constraints*, which helped develop the two major projects contained in this thesis. I have come to appreciate that finding the right constraints is the main ingredient in correctly identifying the problem to solve, as well as the essential perspective to accurately model a behaving system. Although it may sound trivial, I find this approach a fundamental lens that allows you to see the core features of a system, and understand its place among its family of systems by studying the possible instantiations of the constraints. In everyday research, the major questions about agency described above are a bit too vague and general, so there were two more specific questions that ultimately were developed during my doctorate thanks to the notion of constraints: How are our decisions and behavior shaped by constraints and how should we model and study them? How do constraints produce an agent's goals and how are they related to its behavioral variability? The first question is tackled in Chapter 2 and 3 by developing models of the Breadth–Depth dilemma. The second question is addressed in Chapter 4, with the proposal of path occupancy maximization as an overarching principle of behavior. Finally, in Chapter 5, I tie the frameworks presented in the Introduction with the works developed during my PhD and show how keeping the big questions in mind can motivate, drive and suggest present and future research.

Contents

List of Figures

Chapter 1

INTRODUCTION

The search for explanations for particular behaviors in the natural world has prompted philosophers, naturalists and scientists to formulate a plethora of theories and frameworks that encapsulate a diverse set of features of behavior, such as mechanism, function, development and evolution of behavioral traits. Ultimately, the explanations that anyone is able to reach are greatly biased and constrained by their own interests. Clearly, when dealing with complex phenomena such as naturalistic behavior, one needs to find the level of analysis that is appropriate to the questions asked (Anderson, 1972), such that the *explanandum* is formulated precisely with respect to the methodology and aims of the scientist's repertoire of reachable *explanans*. This is the main reason why there is a vast amount of fields that study “natural" behavior, as they do so from a diverse set of methodologies and aims. Additionally, just like the behavioral repertoires of an organism are constrained by its phylogeny and development, scientific fields are constrained by the evolution of their own predecessors and of the ideas around the subject of study. (There is a reason why the phrase popularized by Isaac Newton is “if I have seen further, it is by standing on the shoulders of giants" (Newton, 1675), and not by standing on their wings or fins.) While these biases are inevitable, it is of utmost importance to take an active role to be aware of the particular aims and methods of the fields to which we subscribe.

In this Introduction, the interests and biases of this thesis about optimal behavior will be delineated as clear as possible by reviewing a necessarily limited but interconnected set of concepts, fields and frameworks that seek to either explain naturalistic behavior or design optimal policies by incorporating two main features of this thesis: optimality and constraints. Specifically, the frameworks reviewed will be Tinbergen's four questions, Marr's three levels of analysis, and, as a case study, Bounded Rationality. They will be reviewed independently first, with a specific link to the works in this thesis, in order to reveal the connections between them in the last section of this Introduction. Chapter 2 presents a bounded optimality model of strategic decision making where sampling constraints are made explicit and where it is shown how the optimality of behavior depends on the capacity of the (abstract) agent, which is itself a function of its sampling resources and the environment statistics. Chapter 3 presents an alternative principle of behavior where (abstract) external reward is not maximized, but reconceptualized as a stimulus that allows for a better occupation of (abstract) action-state path space. Chapter 4 discusses the contributions of these works and unifies them with the core idea of constraint as a powerful ingredient for obtaining insights about agency and optimality, as well as the potential applications and future directions of the works presented in this thesis.

1.1 Tinbergen's four questions

In Niko Tinbergen's 1963 landmark article (Tinbergen, 1963), he proposed a framework of four "major problems of Biology", for three of which he credited Julian Huxley (Huxley, 2009, p. 40) and then added a fourth. The problems are, in his own words, "that of causation, that of survival value and that of evolution – to which I should like to add a fourth, that of ontogeny" (Tinbergen, 1963). Although matters of causation and function in science and philosophy go back at least as far as Aristotle's four causes (Juarrero, 2000; Hladky and Havlicek, 2013), Tinbergen's four questions were originally specifically devised to understand biological phenomena and, more precisely, behavior. In fact, he advocated the conception of a new field, "The Biology of Behavior", which overlapped the many fields of Ethology, but with the aim of integrating them for a holistic understanding of behavior (Tinbergen, 1963). It is this unifying feature of the framework that is crucial for its importance in the field of Modern Ethology, but which has largely been cast aside, and the individual questions have mainly continued to be tackled separately (Bateson and Laland, 2013).

The four questions have had a big influence on Modern Ethology, although Bateson and Laland (2013) call for a revisit on the taxonomy. They can be read in their original conception in (Tinbergen, 1963) and a modern interpretation in (Bateson and Laland, 2013), but here they are summarized succintly, and are as follows:

- Causation, understood as "mechanism of control", which aims to explain the mechanistic/physiological basis of behavior.

- Survival value, understood as "current utility" or "function", which aims to understand the contribution of the studied behavior to the organism's survival in the present day.

- Evolution, which aims to understand the part that constraints brought by the organism's phylogeny play in shaping the studied behavior.

- Ontogeny, which aims to understand how the behavior came to be in the particular organism, i.e. how the study of the organism's development can provide explanations about the existence and persistence of a particular behavior.

These questions can be subsequently categorized as two 'how' questions (Mechanism and Ontogeny) and two 'why' questions (Current utility and Evolution), also known as proximate versus ultimate view, respectively. Another useful classification of these questions is in terms of the timescale involved; the mechanism and the current utility look for *explanans* in the present condition of the organism whereas the evolution and ontogeny of behavior extend backwards in time. These two classifications are heuristic in nature, as the borders between why and how questions, as well as between present and past, are not universally determined. However, both the proximate-ultimate and the present-past axes covered by these questions help us understand the nature of the explanations one can seek to formulate and, ideally, try to unify. These four questions do not exist without debate about them (Cuthill, 2005), but they have withstood the test of time and continue to prove useful for a wide range of scientific fields (Bateson and Laland, 2013).

All the questions are important for the understanding of any one complex system, and in this thesis we will be concerned with two abstractions that project somewhat to all of them: optimality and constraint.

Optimality

First, all throughout the text, we will be defining **utility functions**, which can be thought as the *purpose* or *function* of a particular agent in a particular situation. In other words, utility functions will be thought as 'why' questions, which have the presumption to give an ultimate explanation of particular actions, be they about current utility or evolved behavioral traits.

Establishing a utility function is necessarily contentious. In the context of Tinbergen's four questions, distinguishing between current utility and evolved traits in particular is important because functions derived from these questions might not be the same. This distinction has been clearly presented by Gould and Vrba (Gould and Vrba, 1982), where they differentiate between an adaptation and an 'exaptation': current function may not equal past function. An example of this is the development of feathers in birds' evolution. According to them, there is sufficient evidence to suggest that feathers were not initially selected for flight, and thus other hypothesis have been put forward, such as for insulation. Then, it was after their appearance that flight was developed, showing them to be an exaptation for flight. On the other hand, while not the original function of feathers, the appareance of flight shaped the continued evolution of the feathered phylogeny, such that the trait was maintained, thus subsequently becoming an adaptation, possibly through many cycles of adaptation and exaptation ranging many generations. It is this entanglement between function and timescale that can gather significant discussion (Cuthill, 2005).

For experimental science, the view is confusingly less controversial: if one wishes to study the behavior of an organism, one trains it to *perform* an externally designed task, and controls for the animal's performance with experimentally tractable interventions. The ability of organisms to show quick adaptations, such as learning, to prepared environments is of paramount importance in any experimental field that observes and studies behavior. However, it is critical to incorporate Tinbergen's questions to enrich the space of reachable explanations.

In any case, once a particular utility function is formulated, in this thesis we will be asking what the sequence of actions that can maximize the utility is, i.e. what is the most suitable behavior with respect to the proposed utility function. As stipulated before, depending on the field, there are many ways to categorize objectives, and there will be an attempt to be as explicit as possible about the nature of the ones introduced in all the work here presented.

Constraints

Second, we will be explicit about **constraints**, particularly in the decision making process. All four questions can be thought of as determinants of constraints, and we will illustrate it with an example: Imagine that you are sitting on a couch, and suddenly feel hungry, so you intend to go to the kitchen for some food.

Current utility constrains the space of possible action sequences and component configurations, such that only the ones that lead to the desired behavior are kept for the current context, which is similar to Juarrero's second order contextual constraints (Juarrero, 2000), and Aristotle's final cause (Hladky and Havlicek, 2013). Getting food responds directly to the hunger signal whose link to the survival value of the organism is clear. This current utility constrains the space of possible behaviors, as one could reach the kitchen in a myriad of ways, but the space- and timescales are restricted by this goal, such that, for example, walking in the opposite direction of the kitchen is largely unfavored.

Evolution constrains the form and function of organisms, in a way that the trajectory throughout evolutionary time impacts the kinds of behavior that organisms can reach, similar to Aristotle's formal cause (Hladky and Havlicek, 2013). Humans are terrestrial and bipedal, such that reaching the kitchen by flying or swimming is impossible, while standing up and walking is the most likely behavioral combination.

Development constrains the form and function of individual organisms by their interaction with their environment. By taking the environment-agent interaction into account, the specific materials and form that an individual can reach are constrained by the developmental process, such as learning and epigenetic phenomena, in a way that mirrors Aristotle's material cause (Hladky and Havlicek, 2013). One particular individual could have learned to stand up by leaning forward and pushing only with their legs, while others could have a preference for pushing with their arms as well. Furthermore, different people walk in different ways and they will decide to prepare different foods adjusting to their possibly diverse preferences carved throughout their lifetimes.

Finally, mechanisms constrain the immediate space of possible paths that lead to a particular behavior through the possible physical realizations of the function. Historically, the search for mechanistic explanations has occupied much of modern science's workforce and, as a result, it has been the main source of answer to the question “what is the cause of that phenomenon?" (Juarrero, 2000). While Tinbergen ambiguously equated the mechanistic question with the word “causation", he advocated this approach to be called simply “The Physiology of Behavior" (Tinbergen, 1963). While the mechanistic pathway for getting food might be extremely complicated throughout scales, one can think that the homeostatic signal of being hungry effects a system-wide response, from particular higher level value-based decisions to possible neuro-muscular activations, which provide material constraints to the food-getting behavior.

In this thesis, we will abstract the various sources of constraints in different ways to simplify, wherever possible, their influence on the kinds of behaviors that agents can implement. We will establish that constraints are crucial to think about ‘rational’ behaviors, given that they determine the accessible action and state space that agents can reach.

1.2 Marr's levels of analysis

During the 1970's, David Marr and Tomaso Poggio independently arrived at but jointly formulated a framework of levels of analysis at which one needs to understand any complex, information-processing system (Marr and Poggio, 1976). The levels were further formalized and popularized in David Marr's 1982 book Vision, and are briefly described in Figure 1-4 (Marr, 2010, p. 25),

> - Computational theory: What is the goal of the computation, and what is the logic of the strategy by which it can be carried out?

- Representation and algorithm: How can this computation be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
- Hardware implementation: How can the representation and algorithm be realized physically?

The original formulation was intended to be general enough to understand any complex system, as Marr elegantly expressed it in his book,

> If one hopes to achieve a full understanding of a system as complicated as a nervous system, a developing embryo, a set of metabolic pathways, a bottle of gas, or even a large computer program, then one must be prepared to contemplate different kinds of explanation at different levels of description *that are linked* (emphasis added), at least in principle, into a cohesive whole, even if linking the levels in complete detail is impractical. (Marr, 2010, p. 20)

However, the specific levels framework proposed by Poggio and Marr (which, from now on, for compactness, will be called "Marr's levels", keeping in mind the original credit to Poggio) was applied by them and has been applied since then with the specific focus of *information-processing* systems.

Optimality

The link between optimality and Marr's levels of analysis is clear from Marr himself,

> It becomes possible, by separating explanations into different levels, to make explicit statements about what is being computed and why and to construct theories stating that what is being computed is optimal in some sense or is guaranteed to function correctly. (Marr, 2010, p. 19)

In other words, he forms a connection between the computational level, where one specifies *what* is being computed and *why*, to a normative account of the system's behavior, where one can define the system's *goal* and give a formal quantification of how well the system is performing a specific task.

In his book, Marr gives specific attention to the computational level of analysis, which he claims had been neglected by the mainstream neurophysiology programs. In the previous decades leading up to the writing of his influential book, much progress in neuronal recordings led to the development of a large reductionist program, where researchers focused on the mechanisms regarding neuronal responses as a function of external stimuli and disregarded ecological, contextual and all other higher level constructs in order to describe the processes of single neurons. It was a reductionist approach simply because it led to the increased attention to the functioning of single neurons *with the goal of understanding the higher order system that they collectively form*, as illustrated by Horace Barlow's first of his five dogmas,

> A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells, and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. **There is nothing else 'looking at' or controlling this activity** (emphasis added), which must therefore provide a basis for understanding how the brain controls behaviour. (Barlow, 1972, p. 380)

Since then, technological advances have further biased the neuroscientific field into looking at the microscopic level, largely in disregard of higher order phenomena, such as behavior, with the hopes that a mechanistic explanation will be sufficient for the scientific endeavor of understanding the complex systems that realize those mechanisms (Krakauer et al., 2017).

The computational level, which seeks to abstract the goal of an information processing system, is crucial to the development of the **utility functions** that are presented in this thesis. By studying formally the optimal solutions for a specific utility function, one can then compare the mathematical analysis to the target behavior to gather insights about the overarching behavioral goal, the algorithms employed or the physical implementations in the organism.

Constraints

The notion of constraint permeates all three levels of analysis in different forms, and in fact specifying each level can be thought of as providing constraints to the neighbouring levels, in both top-down and bottom-up directions. To illustrate the constraints imposed by the different levels, let us use the same example as in the previous section: being on a couch, feeling hungry and getting food at the kitchen. For practical purposes, one can imagine the individual to be a robot that we are programming to perform this task. First, let us work in a top-down fashion. The computational level is concerned with the 'what' and 'why' questions. The 'what' question describes the computation done by the system, thus constraining the space of possible tasks. More importantly, the 'why' question can be thought of as specifying constraints about why certain computations are more appropriate for the goal, as opposed to others. For the food-getting example, one can identify the computational problem to be about locomotion: the goal in that particular context is getting food by physically moving body parts to reach the kitchen and prepare food. To do so, complex action-perception feedback loops will likely have to be used. The reason why locomotion is the computational problem comes from the requirements or *constraints* given by the context, i.e. the individual (robot) is hungry and needs to displace themselves to another physical location. The usefulness of conceptualizing the computational level as specifying constraints was pointed out by Marr in his book,

> [...] the business of isolating constraints that are both powerful enough to allow a process to be defined and generally true of the world is a central theme of our inquiry. (Marr, 2010, p. 23)

In a different sense, the computational level will also interact with the algorithmic level by constraining the space of possible representations and algorithms to be implemented. By identifying the computational level as locomotion, the representations will have to somehow include the right level of relevant variables, such as body positions, velocity of the limbs or the acceleration of the center of mass. In fact, there might be a wide range of choices for the specific representations, such as position with respect to the kitchen versus the center of mass, egocentric versus allocentric visually-guided navigation, etc. There might also be a variety of algorithms, working with the chosen representations, that instantiate the computational problem, such as leaning forward to get up versus jumping off the couch, walking versus running to the kitchen, etc.

While the computational level constrains the space of possible representations and algorithms, the algorithmic level will itself constrain the space of possible physical instantiations. If, for example, the chosen algorithm involved slowly getting up and walking to the kitchen, all the neuromuscular activations that are associated with jerky, imprecise movements of the limbs when jumping off the couch and running cannot be utilized.

On the other hand, this constraint cross-talk between levels of analysis can also be applied in the reverse direction, that is, bottom-up. If we start off with the particular materials that the individual's body is made off, this itself will constrain the types of representations and algorithms that can be expressed. For example, a human body, with its computational brain and particular anatomy navigates any physical space in a much different way than a robot or a bird would. In fact, the actual algorithmic operations involved in any navigational task are somehow already embodied and have been shaped by evolution: we as humans have two frontal eyes, a central and peripheral nervous system, knees that bend outwards, etc. If a bird is confronted by the same task (for example, being on a couch and going to a kitchen to get food), its reachable representations and algorithms will be radically different to a human's. In a parallel way, the particular representations and algorithms that are reachable by the hardware also constrain the types of tasks, goals and computations that any information-processing system will be able to do. For example, while the human anatomy and processing power allow for flexibility in navigation by walking, running or tumbling, flying without any aid is out of the question, such that reaching a floating kitchen in the air is not an allowed task for humans.

The interaction between levels is not something that Marr originally intended to investigate. In fact, in his book, while he conceded that the links between the levels are crucial; he advocated that they should be studied independently (Marr, 2010). Since then, more attention to the specific links between levels has been given, particularly under the lens of constraints. For example, a perspective that takes into account the specific bottom-up constraints from the implementational level to the algorithmic level all the way to the computational level is given by the efficient coding hypothesis, which began as an application to Claude Shannon's information-theoretic ideas to the information processing of sensory neurons (Attneave, 1954; Barlow, 1961). In it, it is hypothesized that (sensory) neurons are functionally and computationally constrained to reflect the specific statistics of the environment that the organism lives in. By implicitly considering the evolutionary, developmental and behavioral trajectory of the organisms, neurons are therefore hypothesized to encode the relevant stimuli *efficiently* (Simoncelli and Olshausen, 2001; Sterling and Laughlin, 2017).

Another example comes from abstracting resource limitations, such as computational constraints, and incorporate them into the algorithmic level. The abstract constraints delineate what types of computations are realizable and, therefore, the goals of the organism are impacted in a way that the optimal behavior is now a function of the available resources (Russell and Wefald, 1991; Russell and Subramanian, 1994; Griffiths et al., 2015). The next section covers Bounded Rationality, a general framework that captures these ideas appropriately for the works later presented in this thesis.

1.3 Bounded rationality

The theoretical and quantitative study of economic behavior was formalized in the influential work of von Neumann and Morgenstern (1953), where they introduced axioms to formalize utility theory and presented the principle of Maximum Expected Utility and a notion of rational behavior, which they define through the following phrase: "the individual who attempts to obtain these respective maxima is also said to act "rationally." But it may safely be stated that there exists, at present, no satisfactory treatment of the question of rational behavior." (von Neumann and Morgenstern, 1953, p. 9). By proposing a set of desired properties about preference orderings, it is possible to relate ideal decision to the actions that maximize an agent's expected utility, which considers the probability of states in the worlds given the action (Gershman et al., 2015). The formalization of this problem made analysis of behavior and, most importantly, derivation of ideal behavior tractable and quantifiable. They knowingly made a normative assumption of complete information, in which the subject or subjects have all the information and computational resources to arrive at the ideal solutions,

> [...] we cannot avoid the assumption that all subjects of the economy under consideration are completely informed about the physical characteristics of the situation in which they operate and are able to perform all statistical, mathematical, etc., operations which this knowledge makes possible. The nature and importance of this assumption has been given extensive attention in the literature and the subject is probably very far from being exhausted. We propose not to enter upon it. The question is too vast and too difficult and we believe that it is best to " divide difficulties." I.e. we wish to avoid this complication which, while interesting in its own right, should be considered separately from our present problem. (von Neumann and Morgenstern, 1953, p. 30)

Sure enough, people realized soon after that predicting human behavior was still out of reach for the field of Economics, whose theories based on rational behavior were not satisfactory, and for the field of Psychology, whose theories based on biases and heuristics were still insufficient. As a consequence, Herbert A. Simon proposed to revise the idea of the "rational man" present in economic theory by considering the decision maker's limitations, thus spawning the field of bounded rationality,

> Broadly stated, the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist. (Simon, 1955, p. 99)

In concrete, Simon proposed to leave the optimization perspective by trying to establish a theory of choice that is apparent from human behavior. Looking at it from Marr's levels, he tried to focus on the algorithmic limitations to explain choice behavior, arguing that the computational problem is not an optimization problem, but a "satisficing" one. In many instances, he claims that this optimization perspective involve a use of unrealistic assumptions, such as a 1) the ability to attach a definite pay-off for each possible outcome, 2) complete ordering of pay-offs, and 3) complete knowledge of the world model. For these three "unrealistic" assumptions, he provides three more realistic situations for which there are “simplifying” solutions based on his satisficing assumption: 1) Simple pay-off functions, where he proposes a discrete set of values, onto which many outcomes can be mapped. This allows for simpler algorithms that can find satisficing states and actions. (2) Consider partial-ordering of pay-offs, where it is unfeasible to compare many features of the alternatives (“oranges to apples”). Then, one can consider an extension of the simplified pay-off function, where he introduces a vectorial pay-off function, and a threshold determined for the value of each feature, a simplification he calls aspiration levels. (3) Incorporating into the decision-making process the information gathering process, thus considering that the sampling of the environment is not costless. By considering this algorithmic limitation, he argued that the simplified pay-off function can help the decision maker perform a more efficient search, reaching a satisfying solution without the extensive use of resources.

Simon’s view about abandoning the notion of optimization is best understood with his own words,

> The question of how it is to behave "rationally," given these limitations, is distinct from the question of how its capabilities could be increased to permit action that would be more "rational" judged from the mountain-top of a more complete model [One might add: “or judged in terms of the survival value of its choice mechanism” (in footnote)]. The two viewpoints are not, of course, completely different, much less antithetical. We have already pointed out that the organism may possess a whole hierarchy of rational mechanisms - that, for example, the aspiration level itself may be subject to an adjustment process that is rational in some dynamic sense. Moreover, in many situations we may be interested in the precise question of whether one decision-making procedure is more rational than another, and to answer this question we will usually have to construct a broader criterion of rationality that encompasses both procedures as approximations. Our whole point is that it is important to make explicit what level we are considering in such a hierarchy of models, and that for many purposes we are interested in models of "limited" rationality rather than models of relatively "global" rationality. (Simon, 1955, p. 16)

This perspective is at odds with the idea of constrained optimization, a formal way to restrict the optimization procedure to follow specified, arbitrary constraints. Gigerenzer and Selten (2000) argue, in all likelihood accurately, that the original ideas of Simon (1955) have been wrongly applied as constrained optimization, which is “inappropriate” and “misleading”. In contrary, they argue that models of bounded rationality use “fast and frugal stopping rules for search that do not involve optimization”(Gigerenzer and Selten, 2000, p. 12).

What is missed from this perspective, however, is the fact that a careful study of constrained optimization can lead to tractable mathematical solutions, where in fact the constrained optimal solution could easily be implemented, thus providing a candidate explanation for the function of a particular computation. The fact that the theorist needs to flex some muscle to arrive at the constrained optimal solutions, does not mean that 1) it is the only way at arriving at the solutions, 2) we think that this is exactly the way the brain or mind do it and 3) evolution and development do not aid in biasing and constraining the space of possible computations. In fact, their idea of heuristic decision making is not incompatible with constrained optimization; the difference being that the former is a descriptive account for behavior, whereas the latter is a prescriptive one. The importance of a prescriptive, normative description was already recognized in the work of von Neumann and Morgenstern (1953), as an assumption of complete information might actually explain some phenomena for which, traditionally, there might have been assumptions of incomplete information, therefore providing a satisfactory interpretation about the agent's state of information (von Neumann and Morgenstern, 1953, p. 30).

The field of Bounded Rationality, understood under this non-optimizing light, then steered towards the observation of actual human behavior to try to find the algorithmic approximations made by humans, illustrated by heuristics, analyzing *a posteriori* why they work. Nevertheless, the idea that rational behavior needed to be revised and formalized to include the cognitive limitations of the decision makers was groundbreaking by itself, prompting many other works to build upon it. The idea has been visited many times, in many forms and with many names since then. Rather than a historical account, we proceed to highlight a few approaches that are relevant to the particular phylogeny of this thesis.

1.3.1 Metalevel rationality

Horvitz (1987) addressed the idea of computational constraints in the context of rational decision making under uncertainty, for the problem of inference in particular. In this work, he analyzed the metareasoning problem (or reasoning about reasoning); in other words, finding the best course of action for computing solutions to apply in the object level, consistent with Good's type II rationality (Russell and Wefald, 1991). The problem was picked up and formalized by Russell and Wefald (1991). They define the object level as the level where actions take place, but the optimization procedure lies in the metalevel, or computational level (not to be confused with Marr's), where the maximization of utility is over the space of *computations*. The conceptual novelty was to introduce the "value of computation", where they explicitly model the cost of computation and weigh it against the utility of the choice, which is a consequence of the selected computation.

One of the main issues, discussed in the paper, is that the metalevel problem is in fact more difficult to solve than the object level, as it has to take into account the space of actions and object-level states, as well as the space of computations.

However, this approach was later picked up by Griffiths et al. (2015) in an attempt to build an analysis of human behavior that takes into account resource limitations. They, following Anderson (1990), made the important observation that Marr's levels make the assumption that the analysis of the computational level can be formulated independently of the details of the algorithmic and physical implementations, which implicitly creates an ideal entity that can solve this problem. This is an adaptationist stance that, when made explicit, prompted Anderson (1990) to create his six-level system of rational analysis,

> 1. Precisely specify what are the goals of the cognitive system.
> 2. Develop a formal model of the environment to which the system is adapted (almost certainly less structured than the standard experimental situation).
> 3. Make the minimal assumptions about computational limitations. This is where one specifies the constraints of evolutionary history. To the extent that these assumptions are minimal, the analysis is powerful.
> 4. Derive the optimal behavioral function given items 1 through 3.
> 5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
> 6. If the predictions are off, iterate. In my own experience, my problems have been with the mathematical analyses required in step 4, which can often be quite complex. (Anderson, 1990, p. 29)

Griffiths et al. (2015) focus on the third one, which summarize their resource-rational analysis,

> Rather than blurring these lines and building constraints into computational-level theories, we suggest a different approach: Define the computational-level theory without considering limitations on its execution, and then explore the consequences of those limitations as a further analysis that brings us closer to an algorithmic-level theory. (Griffiths et al., 2015)

Resource-rational analysis is then a process of finding the algorithms that make optimal use of the cognitive constraints at each level, until hopefully eventually finding the right ones. The focus is then on idealizing a computational level (that can be solved by an unbounded agent) and thinking that real behavior has an algorithmic capability with finite resources. By explicitly modeling computation costs, they can find the optimal use of the assumed resources for each particular family of algorithms, and then compare that to real behavior. They claim that the process of finding the "correct" approximation that best matches real behavior is valuable. That is, the quest is to find insights into the algorithmic constraints and processes by assuming that real agents have the capability to implement the theorist's family of algorithms. Crucially, they augment the utility function to include the cost of computation, which then will favor (or disfavor) particular algorithms that will try to maximize utility. It is thus essentially an adaptation of rational metareasoning introduced by Russell and Wefald (1990), where one introduces a metalevel problem where the optimization procedure is defined in the space of computations (or algorithms), constrained by the cost of those computations.

1.3.2 Bounded optimality

Horvitz (1987) also coined the term 'bounded optimality' to "refer to the optimization of computational utility given a set of assumptions about expected problems and constraints in reasoning resources." (Horvitz, 1987). However, the problem was not properly formalized and the term changed its meaning thanks to Russell and Subramanian (1994)'s work on provably bounded optimal agents. In this article, they borrowed the term introduced by Horvitz and formalized it mathematically. Consistently with Horvitz's definition, they define a bounded optimal agent as one that "behaves as well as possible given its computational resources. Bounded optimality specifies optimal programs rather than optimal actions or optimal computation sequences."(Russell and Subramanian, 1994) The motivation for revisiting this concept stems from the fact that maximizing over the space of actions or computations makes specific, unrealistic assumptions about the implementability of those actions and computation sequences by the agents. Metareasoning deals with the fact that action sequences and their consequences need to be computed with limited resources, which changes the notion of optimality at the object level. However, those optimal metapolicies themselves need resources to be computed, and one enters an infinite regress that cannot guarantee that an agent will be able to implement the (resource rational) optimal object-level policy function. One way out of this infinite regress is to introduce the idea that we can model any decision-making agent as a machine architecture that runs programs, whose space is defined by a programming language and is therefore necessarily finite. One defines a utility function in the same way as always, but instead of maximizing it over the space of actions, computations (or computations of computations, so on and so forth), the idea is to find the program that maximizes this utility, which can always be done (Russell and Subramanian, 1994). This approach then puts the emphasis on the architecture of the agent, instead of on the policy functions, which are actually realized by programs running on the architecture.

This idea was picked up by Lewis et al. (2014) to apply it as a framework of analysis in cognitive science. The approach to analyzing behavior under the lens of bounded optimality was named as 'computational rationality'. This approach puts special emphasis on the fact that behaviors are generated by cognitive architectures that are adapted not only to the structure of environments they have evolved in, but also to the structure of "the mind itself". When behaviors can be better explained by introducing both, Lewis et al. (2014) call this type of rationality “ecological-bounded optimality”. This channels one of Herbert Simon’s first works on bounded rationality, whose main effort was to show that a bounded rational agent should exploit ecological regularities to achieve their satisficing needs,

> [...] if an organism is confronted with the problem of behaving approximately rationally, or adaptively, in a particular environment, the kinds of simplifications that are suitable may depend not only on the characteristics – sensory, neural, and other – of the organism, but equally upon the structure of the environment.(Simon, 1956)

As a contrast to resource-rational analysis, computational rationality does not interpret behavior as approximations of unbounded optimality, but rather that behavior is a *consequence* of the agent adapting to both the environment and its own cognitive and physical architecture. The applicability and usefulness of resource-rational analysis versus computational rationality is a matter of perspective, and ultimately it will depend on the ease of use of their proposed frameworks. The former focuses on the space of approximations induced by resource constraints to a global optimization, whereas the latter focuses on the space of optimal programs and machines that are a consequence of an ecological-mechanistic adaptation, which induces by itself a notion of resource constraints.

1.4 Discussion

Thus far, we have presented two frameworks for the analysis of behavior: Tinbergen’s Four Questions, Marr’s levels of analysis, and one field that makes specific assumptions about the real cognitive structures that produce behavior, Bounded Rationality. They were all originally formulated for distinct purposes and fields, but share some properties that are highly relevant for the development of this thesis. The idea of Tinbergen’s Four Questions and Marr’s levels is to provide any theory with a pluralistic view of the studied phenomenon and therefore have specific important value to the understanding of behavior, which has been pointed out before for the specific field of Neuroscience (Krakauer et al., 2017). In this thesis, a similar view is advocated. In particular, the existence of these frameworks lets us study our theories in a structured way, and gives us a guide into possible gaps in our understanding. In fact, we can use them to study the structure of the frameworks proposed under the Bounded Rationality umbrella.

As we have reviewed, Bounded Rationality posits that we need to incorporate the specific implementational limitations of agents in our theories to understand their behavior, and we can use Tinbergen's Four Questions to analyze it. First, it is straightforward to see that the function of a behavior asks what the utility of a particular behavior is in the *current* environment, providing a static *why* of the behavior, which would correspond directly to the utility function or aspirations in Bounded Rationality. Secondly, it is important to recognize that the mechanisms that lead to the behavior have a deep history of adaptation to a family of environments. The constraints brought by evolution need to be taken into account to understand a dynamic *why* of the behavior. This question maps onto the ecological part of ecological-bounded optimality, as well as it provides great motivation to understand the origin of the resource limitations of Griffiths et al. (2015) or the architectures of Lewis et al. (2014). Thirdly, the actual environment where the agent is in has a great influence on the behavioral mechanisms it can display via its development, which provides the theorist with and understanding of *how* the resource constraints are formed, and how they are continuously impacted by the current environment through learning. Agents that are resource limited develop in a particular environment are encouraged to learn to exploit its structure regularities, which maps onto the original idea of Simon (1956). Finally, understanding the physical and cognitive machinery of agents will provide clues about *how* they can produce behavior and actually realize the function, which maps onto the actual study of the resource limitations in Bounded Rationality.

Parallely, we can use Marr's levels of analysis to probe the structure of the Bounded Rationality theories and frameworks, distinguish between them and extract different dimensions of understanding from them. In the case of the first level, Bounded Rationality, as envisioned by Simon (1955) and advocated by Gigerenzer and Selten (2000), does not define the computational level as being about optimizing a utility function, but rather to "satisfice" it. In contrast, bounded optimality Russell and Subramanian (1994) and metareasoning Russell and Wefald (1991) frameworks do involve a constrained optimization procedure, although they make different assumptions about the utility function. The major conceptual work in this field lies on the second level. It is precisely by focusing on the algorithmic limitations of humans that a bounded-rational analysis of their behaviors was put forward by Simon (1955). However, the different approaches that stemmed from it differ in how they conceptualize the resource limitations. For example, metareasoning and resource-rational analysis consider resource limitations as costly computations, whereas bounded optimality and computational rationality consider resource limitations as having a finite set of programs that can run on the particular architecture of agents. Finally, the physical implementations of the algorithms are not usually a subject of bounded rationality accounts, although the fact that physical substrates have to do the actual computing definitely informs the algorithmic level, as seen in the field of efficient coding (Barlow, 1972; Simoncelli and Olshausen, 2001; Sterling and Laughlin, 2017). Importantly, many of the accounts that span the bounded rationality spectrum advocate, in their own ways, using the algorithmic level to inform the computational one. This integration of levels was something that Marr (2010) called for, even if originally he envisioned their independent analysis.

In Chapter 2 and 3, we will present a model of strategic decision making that is motivated by the Bounded Rationality field. In brief, we explore the optimal allocation of limited resources in a situation where feedback is delayed, and thus the optimal allocation needs to take both the environment regularities into account and the actual capacity of the agent to sample the environment, in line with Simon's scissor-like view of human cognition (Simon, 1956; Gigerenzer and Selten, 2000), as well as with the ecological-bounded rationality proposed in Lewis et al. (2014).

The fact that there is extensive discussion about the nature of rationality of agents, especially in humans, given their behaviors, and thus looking for explanations for it in terms of perfect rationality (von Neumann and Morgenstern, 1953), metalevel rationality (Russell and Wefald, 1991), bounded optimality (Russell and Subramanian, 1994), resource-rational analysis (Griffiths et al., 2015), ecological rationality (Anderson, 1990), ecologically bounded optimality (Lewis et al., 2014) or any other framework to analyze behavior, makes it apparent that it is quite hard to (1) determine the function of a behavior and (2) attach the relevant computational constraints to it. Therefore, inferring a utility function from behavior or designing one for artificial agents is a problem that arises by establishing the need to associate behavior to utility maximization.

An alternative formulation is to endow the agents with a general enough principle of behavior, and cast function, utility or reward as constraints for this principle. For example, if we establish that an overarching principle of behavior is for agents to survive, then particular physical substrates and dynamics, learning and developmental mechanisms, ecologies, evolutionary histories and cognitive capabilities all provide *constraints* for this principle to be achieved. Whether this is a principle that predicts behavior accurately or not is an empirical question. Other alternatives can be tried, such as reproduction or group survival, in order to capture real behavior. In any case, what this perspective provides is that the usual utility functions, such as external reward, and the associated intervening processes that are studied in single tasks, such as algorithms, representations and physical implementations, can now be reinterpreted as serving a higher level process. In Chapter 4 we will propose one such principle of behavior to reconceptualize core ideas in modern research such as reward, curiosity and exploration.

Finally, in Chapter 5 we will discuss the contributions of the works, and contextualize them into the frameworks presented in this Introduction, to subsequently visit the possible promising scientific paths to take in the future given the ideas here presented.

Chapter 2

BREADTH-DEPTH DILEMMA

The following chapter is based on the manuscript published in Proceedings of the National Academy of Sciences, see (Moreno-Bote et al., 2020) for bibliographic details. I declare to be one of the main authors of this work.

Abstract

In multi-alternative risky choice we are often faced with the opportunity to allocate our limited information-gathering capacity between several options before receiving feedback. In such cases, we face a natural tradeoff between *breadth* – spreading our capacity across many options – and *depth* – gaining more information about a smaller number of options. Despite its broad relevance to daily life, including in many naturalistic foraging situations, the optimal strategy in the breadth-depth tradeoff has not been delineated. Here, we formalize the breadth-depth dilemma through a finite sample capacity model. We find that, if capacity is small (around 10 samples), it is optimal to draw one sample per alternative, favoring breadth. However, for larger capacities, a sharp transition is observed, and it becomes best to deeply sample a very small fraction of alternatives, that roughly decreases with the square root of capacity. Thus, ignoring most options, even when capacity is large enough to shallowly sample all of them, is a signature of optimal behavior. Our results also provide a rich casuistic for metareasoning in multi-alternative decisions with bounded capacity using close-to-optimal heuristics.

2.1 Introduction

The breadth-depth (BD) dilemma is a ubiquitous problem in decision-making. Consider the example of going to graduate school, where one can enroll in many courses in many topics. Let us assume that the goal is to determine the single area of research that is most likely to result in an important discovery. One cannot know, even in a few weeks of enrollment, whether a course is the most exciting one. Should I enroll in few courses in many topics –breadth search— at the risk of not learning enough about any topic to tell which one is the best? Or should I enroll in many courses in very few topics –depth search— at the risk of not even taking the course with the really exciting topic for the future? One crucial element of this type of decision is that the resources (time, in this case) need to be allocated in advance, before feedback is received (before classes start). Also, once decided, the strategy cannot be changed on the fly, as doing so would be very costly.

The BD dilemma is important in tree search algorithms (Horowitz and Sahni, 1978; Korf, 1985) and in optimizing menu designs (Miller, 1981). It is also one faced by humans and other foragers in many situations, such as when we plan, schedule, or invest with finite resources while lacking immediate feedback. Furthermore, it is a dilemma that a large number of distributed decision-making systems have to tackle. These include, for example, ant scouts searching for a new colony settlement (Pratt et al., 2002), stock market investors, or soldiers in an army during battle. Evidence suggests that distributed processing with limited resources is also a valid model of brain computations (Balasubramani et al., 2018; Eisenreich et al., 2017). In face of this, it is remarkable that the bulk of research on the BD has been in fields outside of psychology and neuroscience (e.g. (Halpert, 1958; Schwartz et al., 2009; Turner et al., 2002)). We believe that one reason for this is the lack of models and formal tools for thinking about the BD dilemma and separating it from other dilemmas.

Many features of the BD dilemma warrant its study in isolation. First, BD decisions are about how to divide finite resources, with the possibility of oversampling specific options and ignoring others, e.g., one can select several courses on the same topic while ignoring other topics. Secondly, the BD dilemma is about making strategic decisions, that is, decisions that need to be planned in advance and cannot be changed on the fly once initiated, e.g., it is very costly to change courses once they have started, at least during the first semester. Finally, BD decisions need to be made before the relevant feedback is received, e.g., enrollment happens before courses start, and thus before knowing the true relevance of the courses and topics. One can easily imagine replacing *courses* by ant scouts or neurons, and *topics* by potential new settlements or sensory functions, and so on, in the above example to reveal new relevant BD dilemmas pertaining to distributed decision making or brain anatomy, respectively.

The identifying features of the BD dilemma are distinct from those of the well-known exploitation-exploration (EE) dilemma (Cohen et al., 2007; Costa et al., 2019; Daw et al., 2006; Ebitz et al., 2018; Wilson et al., 2014) and its associated formalization in multi-armed bandits (Averbeck, 2015; Chen et al., 2016; Gittins et al., 2011). Specifically, whereas in the EE dilemma samples are allocated sequentially, one by one, to gather information and reward after each sample, in the BD dilemma multiple samples can be allocated in parallel at once to multiple options (possibly allocating multiple samples to some) without immediate feedback to gather information and maximize future reward. It is worth pointing out that EE and BD are not mutually exclusive aspects of decision making, and therefore they are expected to appear hand-in-hand in many realistic situations.

Past work has revealed that humans appear to carefully trade off the benefits of examining many options broadly and examining a smaller number of options deeply in multi-alternative choice. For example, when faced with a large number of options, we often focus – even if arbitrarily – on a subset of them (Bettman et al., 1998; Brandstätter et al., 2006; Gigerenzer and Gaissmaier, 2011; Tversky, 1972) with the presumable benefit that we can more precisely evaluate them. Likewise, we may consider all options, but arbitrarily reject value-relevant dimensions (Busemeyer et al., 2019; Timmermans, 1993), as if contemplating them all is too costly. Option narrowing appears to be a very general pattern, one that is shared with both human and non-human animals, despite the fact that rejecting options can reduce experienced utility (Gigerenzer and Gaissmaier, 2011; Tversky, 1972). It is often proposed that such heuristics reflect bounded rationality (Simon, 1955), which is likely correct in principle, but the exact processes underlying that boundedness remain to be identified. Why do we so often consider a very small number of options when considering more would a priori improve our choice? One possibility is that this pattern reflects an evolved response to an empirical fact: that when capacity is constrained, optimal search favors consideration of a small number of options.

Because cognitive capacity is limited in many ways, the BD dilemma has direct relevance to many aspects of cognition as well. For example, executive control is thought to be limited in capacity, such that control needs to be allocated strategically (Hills et al., 2010; Koechlin and Summerfield, 2007; Shenhav et al., 2013, 2017). Likewise, attentional focus and working memory capacity are limited, such that, during search, we often foveate only a single target or hold a few items in memory (Cowan et al., 2005). Although the effective numbers are low, each contemplated option is encoded with great detail (Awh et al., 2007; Luck and Vogel, 2013; Ma et al., 2014). Furthermore, it seems clear that recollection of information from memory can be thought of as a search-like process (Hills et al., 2012; Ratcliff and Murdock, 1976; Shadlen and Shohamy, 2016). That is, to retrieve a memory we must attend to a recollection processes, with its associated limited capacity. Thus memory-guided decisions presumably involve BD tradeoffs too.

Although the relevance of the BD dilemma is clear, tractable models are lacking, and thus, optimal strategies for BD decisions are largely unknown. Here, we develop and solve a model for multi-alternative decision making endowed with the prototypical ingredients of the BD dilemma. Our model consists of a reward-optimizing yet bounded decision-maker (Gershman et al., 2015; Simon, 1955) confronted with multiple alternatives with unknown subjective values. The first critical element of the model is *finite sample capacity*, which enforces a tradeoff between sampling many options with few samples each (breadth) and sampling few options with many samples each (depth). The second critical element is that samples need to be allocated across alternatives before sampling starts and, thus, before feedback is available. This strategic decision with the finite sample capacity constraint implies a metareasoning problem (Griffiths et al., 2015; Russell and Wefald, 1991) where deliberation about the multiple possible allocations of resources (meta-actions) need to be made in advance to optimize expected utility of a future choice.

Despite the simplicity of the model, it features non-trivial behaviors, which are characterized analytically. When capacity is low (less than 4-10 samples can be probed), it is best to sample as many alternatives as possible, but only once each; that is, breadth search is favored. At larger capacities, there is a qualitative and sharp change of behavior (a ‘phase transition’) and the optimal number of sampled alternatives roughly grows with the square root of sample capacity (‘square root sampling law’), balancing breadth and depth. Therefore, in the high capacity regime it is best to ignore the vast majority of potentially accessible options. We considered globally optimal allocations in comparison to even allocation of samples across sampled alternatives and found that the square root sampling law, obtained for the latter, provides a close-to-optimal heuristic that is simpler to implement. We also study limit cases where the above rules break down, as well as generalizations to dynamic allocation of finite resources with feedback that illustrate the generality of the results. Our results are also robust to strong variations of the environments where the probability of finding good options widely varies.

2.2 Results

2.2.1 Finite sample capacity model

We assume that a decision-maker can choose how to allocate a finite resource among options of unknown status to determine the best option (Fig. 2.1). The environment generates a large number of options, each characterized by the probability of delivering a successful outcome. The success probabilities, unknown to the decision-maker, determine the quality of each of the options, with better options having higher success probabilities (e.g., options with a higher probability of delivering a large reward if they are sampled). The goal of the decision-maker is to infer which of the options has the highest success probability (and thus, highest expected value). The success probabilities of the options are generated randomly from an underlying prior probability distribution, modelled as a beta distribution with parameters (α, β). We assume that this distribution is known by the decision-maker, due, for example, to previous experience with the environment. The prior distribution determines the overall difficulty of finding successful options in the environment.

The decision-maker is endowed with a finite sample capacity C, i.e., a finite number of samples that she can allocate to any option and to as many options as desired. Within the allowed flexibility, it is possible that the decision-maker decides to oversample some options by allocating more than one sample to them, and it is also possible that she decides to ignore some options by not sampling them at all. Feedback is not provided at the allocation stage, so this decision is based purely on the expected quality of options in the environment. After allocation has been determined, the outcomes of the samples are revealed, constituting the only feedback that the decision-maker receives about the fitness of her sample allocation. Outcomes for each of the sampled alternatives are modelled as a Bernoulli variable, where a successful outcome (corresponding to a large reward) has probability equal to the success probability of that option (see below for a generalization in which we consider Gaussian outcomes). The inferred best alternative is the one with the largest inferred success probability based on the observed outcomes from the allocated samples to each of the options (Bechhofer and Kulkarni, 1984; Gupta and Liang, 1989; Sobel and Huyett, 1957). Choosing this alternative maximizes expected utility (see below and Appendix).

While making a choice based on the observed outcomes is a trivial problem, deciding how to allocate samples over the options to maximize expected future reward is a hard combinatorial problem. There are many ways a finite number of samples can be allocated amongst a very large number of alternatives. At the *breadth* extreme, one can split capacity to over as many alternatives as possible, sampling each just once. In this case, the decision-maker will likely identify a few promising options, but will lack the information for choosing well between them. At the *depth* extreme, the search could allocate all samples only to a couple of alternatives. The decision-maker's estimate of the success probability of those options will be accurate, but that of the other alternatives will remain unknown. It would seem that an intermediate strategy is better than either extreme. Specifically, the optimal allocation of samples should balance the diminishing marginal gains of sampling a new alternative and those of drawing an additional sample from an already sampled alternative.

To formalize the above model, let us assume that the decision-maker can sample and choose from $N = C$ alternatives. That is, we consider scenarios where the number of alternatives N is as large as the decision-maker's sampling capacity –if the number of alternatives is larger than capacity, the only difference is that there would be a larger number of ignored alternatives. The allocation of samples over

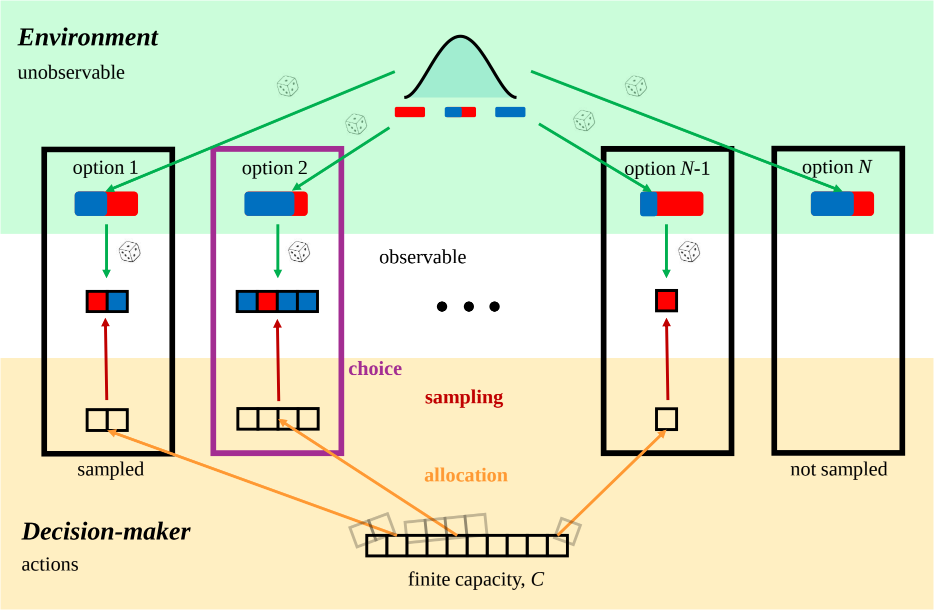


Figure 2.1: Finite sample capacity model. The environment (top, green) contains a large number N of options, and choosing either might lead to a successful outcome (e.g., a large vs a small reward). For each option, the probability of success (blue fraction of red/blue bar) is a priori unknown to the decision-maker and is drawn independently across options from an underlying prior probability distribution, modelled as a beta distribution (top distribution). Options are characterized by the probability of delivering a successful outcome (e.g., a large reward), and the outcomes are modelled as Bernoulli variables. The decision-maker (bottom, orange) has a finite capacity C, i.e., a finite number of samples (bar of squares) that can be allocated to any option in any possible way. All samples need to be allocated in advance and allocation cannot be changed thereafter. Therefore, feedback is not provided at this stage. After allocation, sampling starts (center, white), in which the decision-maker observes a number of successes and failures for each of the sampled options (colored squares; blue: success –large reward, red: failure –small reward). Once this evidence is collected, the decision-maker chooses the option that is deemed to have the highest probability of success (in this case, option 2; purple box).

the alternatives is described by the vector $\overrightarrow{L}$, with components L_i representing the number of samples allocated to alternative $i = 1, \ldots, N$. The finite sample capacity of the decision-maker imposes the constraint $\sum_i L_i = C$. Upon drawing L_i samples from each alternative i, the decision maker observes the number of successes (1's), denoted n_i, of each of the Bernoulli variables. The best option is then the one with the highest posterior mean probability $E\left(p_i|n_i, \alpha, \beta\right) = \frac{n_i+\alpha}{L_i+\alpha+\beta}$ after observing these successes, such that the utility for a given allocation $\overrightarrow{L}$ and associated outcomes $\overrightarrow{n}$ becomes $U\left(\overrightarrow{n}, \overrightarrow{L}\right) = \max_i \frac{n_i+\alpha}{L_i+\alpha+\beta}$. Because the number of successes is only revealed after selecting the sample allocation strategy $\overrightarrow{L}$, the decision-maker's utility for using that strategy, $U\left(\overrightarrow{L}\right)$, is an average of $U\left(\overrightarrow{n}, \overrightarrow{L}\right)$ over all possible outcomes given $\overrightarrow{L}$,

$$U\left(\overrightarrow{L}\right) = \sum_{\overrightarrow{n}} p\left(\overrightarrow{n}\middle|\overrightarrow{L}, \alpha, \beta\right) \max_i \frac{n_i + \alpha}{L_i + \alpha + \beta}, \tag{2.1}$$

where $p\left(\overrightarrow{n}\middle|\overrightarrow{L}, \alpha, \beta\right)$ is the joint probability distribution of the outcomes $\overrightarrow{n}$ given the allocation $\overrightarrow{L}$ and the prior distribution parameters. As each alternative is sampled independently, the distribution of success counts factorizes as $p\left(\overrightarrow{n}\middle|\overrightarrow{L}, \alpha, \beta\right) = \prod_i p(n_i|L_i, \alpha, \beta)$, where $p(n_i|L_i, \alpha, \beta)$ is a beta-binomial distribution (Murphy, 2012). This distribution specifies the probability of observing exactly n_i successes from a Bernoulli variable that is drawn L_i times, and whose success probability p_i follows a beta distribution with parameters α and β. These two parameters control the skewness of the distribution: if both parameters are equal, the distribution is symmetric around one half, while for α larger (smaller) than β the distribution is negatively (positively) skewed.

Finally, the optimal allocation of samples across options $\overrightarrow{L}^*$ is the one that maximizes the decision-maker's expected utility $U\left(\overrightarrow{L}\right)$ in Eq. (1) over all allocations of samples $\overrightarrow{L}$,

$$\overrightarrow{L}^* = \arg\max_{\overrightarrow{L}} U\left(\overrightarrow{L}\right) \tag{2.2}$$

with the above finite sample capacity constraint (see Methods for details). The optimal expected utility then becomes $U^* = \max_{\overrightarrow{L}} U\left(\overrightarrow{L}\right)$, which involves a double maximization over the expected success probabilities of the sampled alternatives and the allocation of samples over the alternatives, effectively solving the two-stage decision process (i.e., first allocate samples, then observe outcomes, and then choose) in reverse order (i.e., first optimize choices given outcomes and allocation, then optimize allocation).

This maximization allows total flexibility over how many samples to allocate to each alternative. However, for the sake of tractability, let us first consider the best even allocation of samples, that is, a subfamily of allocation strategies where the same number of samples L are allocated to each of M sampled alternatives, while the remaining alternatives $(C-M)$ are not sampled, subject to the standard capacity constraint $M \times L = C$. Indeed, finding the optimal even allocation of samples is easier than finding the globally optimal allocation, which might be uneven in general (see below). As we show in Methods, a particularly simple expression for the optimal even sample allocation, L^*, arises when the prior distribution over success probabilities is uniform ($\alpha = \beta = 1$),

$$L^* = \arg\min_L \frac{\sum_{s=0}^{L}(s+1)^M}{(L+1)^M(L+2)}, \tag{2.3}$$

where the right-hand side is related to utility by

$$U(M = C/L) = 1 - \frac{\sum_{s=0}^{L}(s+1)^M}{(L+1)^M(L+2)}. \tag{2.4}$$

Note that only $M^* = C/L^* \leq C$ alternatives are sampled in the optimal allocation, while the remaining options are given zero samples, thus effectively being ignored. The sampled alternatives can be chosen randomly, as they are indistinguishable before sampling. Using extreme value theory (see Methods), we show that the optimal number of sampled alternatives M^* and optimal number of samples per alternative L^* both follow a power law with exponent ½ for large capacity C

$$\lim_{C\to\infty} M^* = \sqrt{C}, \qquad \lim_{C\to\infty} L^* = \sqrt{C}, \tag{2.5}$$

which corresponds to perfectly balancing breadth and depth.

In the next section, we analyze this case in detail. After that, we consider optimal even allocations of samples for arbitrary prior distributions, and finally we provide results for the globally optimal allocations, not necessarily even.

2.2.2 Sharp transition of optimal sampling strategy at low capacity

We first analyze the expected utility $U(M)$ as a function of the number of evenly sampled alternatives M, each sampled L times (such that $M \times L = C$) (Fig. 2.2a) At low capacity ($C = 4$, lighter gray line), the utility increases monotonically from sampling just one alternative ($M = 1$) four times, to sampling four alternatives ($M = 4$) one time each. Thus, a pure breadth strategy is favored. At intermediate capacity ($C = 10$, medium gray line), the maximum occurs at an intermediate number of alternatives (specifically, $M = 5$), reflecting an increasing emphasis on depth. At large capacity ($C = 100$, black line), the maximum expected utility occurs when sampling few different alternatives ($M = 10$ sampled alternatives with $L = 10$ samples each), reflecting a tight balance between breadth and depth. For such large capacities, a breadth search that samples most of the alternatives (rightmost point of the black line) would lead to a reward that approaches 2/3, which is the lowest expected reward one would obtain if at least one sampled alternative has a positive outcome (see Appendix).

The model displays a sharp transition when capacity crosses the critical value of 5 (Fig. 2.2b) Below this critical capacity, the optimal number of sampled alternatives equals capacity. That is, one should follow a breadth strategy and distribute one sample to each alternative. Above 5, the optimal number of sampled alternatives is much smaller than the capacity, with the temporary exception of capacity equal to 7. That is, one should balance the number of sampled alternatives with the depth of sampling each of them. Specifically, the optimal number of sampled alternatives follows a power law with exponent $1/2$ (log-log linear regression, power = slope = 0.49, 95% CI = [0.48, 0.50]), as predicted by Eq. (5), which implies that the fraction of sampled alternatives decreases with the square root of capacity. This means that breadth and depth are tightly balanced in the optimal strategy. The sharp transition at around 5 becomes clearer when plotting the ratio between the optimal number of sampled alternatives and capacity as a function of capacity (Fig. 2.2c).

In summary, if the capacity of a decision-maker increases by a factor of 100, the decision-maker will roughly increase the number of samples alternatives just by a factor of 10, one order of magnitude smaller than the capacity increase. Because the optimal number of sampled alternatives increases with capacity with an exponent ½ , we call this the ‘square root sampling law’. A remarkable implication of this law is that the vast majority of potentially accessible alternatives should be ignored (e.g., for $C = 100$, $C - M = 90$ options are ‘rationally’ ignored).

2.2.3 Generalizing to variations in beta prior distributions

The above critical capacity for optimal even sample allocation changes when, instead of using a uniform prior of success probabilities, we allow for variations of the prior distribution (Fig. 2.2d). However, the critical capacity consistently lies again at around low values (around 10) with the specific value depending on the environment. By changing the prior’s parameters, we can vary the difficulty of finding a good extreme alternative, and thus can compare different scenarios. For the uniform prior that we have used previously (a ‘flat’ environment), a decision-maker is equally likely to find an alternative with any success probability. Consider a prior distribution that is concentrated and symmetric around a success probability

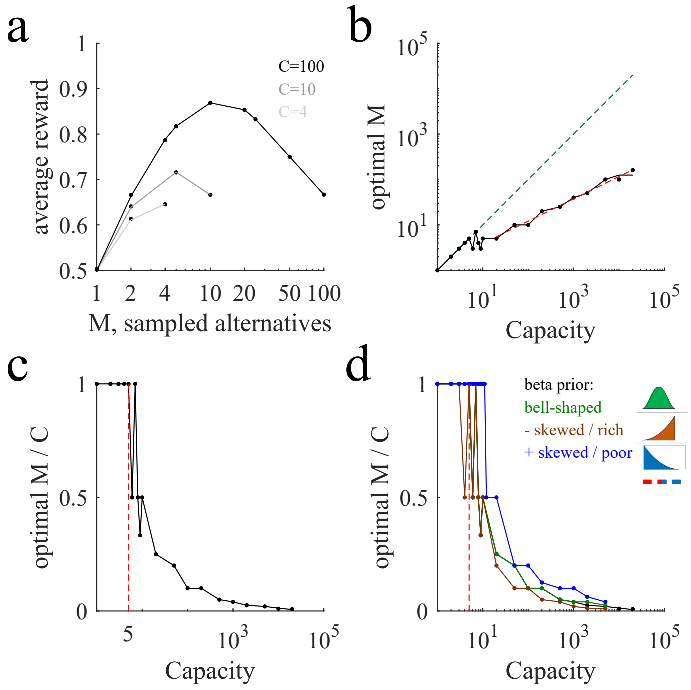


Figure 2.2: Sharp transitions in optimal number of sampled alternatives at low capacity, and power law behavior at large capacity. (a) Average reward (points, simulations; lines, theoretical expressions, Eq. 4) as a function of the number of sampled alternatives M for three different capacities ($C = 4, 10, 100$; light, intermediate and dark lines respectively) for the flat environment (uniform prior). The maximum occurs at the large extreme for low capacity but at a relatively low value for large capacity. Note log horizontal scale. (b) Optimal number of sampled alternatives as a function of capacity. When capacity is smaller than around 5, a linear trend of unit slope is observed (dashed green line), but when capacity is above 7, a sublinear behavior is observed (dashed red line corresponds to the best power law fit, with exponent close to $1/2$). Black line corresponds to analytical predictions. The jagged nature of this prediction and simulation lines in this and other panels is due to the discrete values that the optimal M can only take, not due to numerical undersampling. (c) The sharp transition is clearer when plotting the optimal number of sampled alternatives to capacity ratio as a function of capacity. For low capacity, the ratio is one, but for large capacity the ratio decreases very rapidly. The last point below which the optimal ratio is always one (critical capacity) corresponds to capacity equal to 5 (indicated with a vertical red line). (d) Number of sampled alternatives to capacity ratios for different prior distributions ($\alpha = \beta = 3$, bell-shaped, green line; $\alpha = 3, \beta = 1$, negatively skewed prior modelling a 'rich' environment, brown line; $\alpha = 1, \beta = 3$, positively skewed distribution modeling looking for a 'needle in a haystack', that is, a 'poor' environment, blue line). Lines correspond to analytical predictions from Eq. 9 in the Methods; points correspond to numerical simulations; error bars are smaller than data points in all panels.

of 0.5 (approximately as a Gaussian, corresponding to the beta prior parameters $\alpha = \beta = 3$). In this environment, unusually good (high success probability) and unusually bad (low success probability) options are rarer than medium ones (Fig. 2.2d, green line). In this case, the breadth-depth tradeoff as a function of C is remarkably similar to the uniform prior case, with a transition at $C = 5$.

We also consider a negatively skewed prior distribution ($\alpha = 3,\ \beta = 1$). This distribution refers to environments with rare bad options, as, for example, a tree whose fruits are mostly ripe but that has a few unripe ones. In this 'rich' environment, one can afford sampling a smaller number of options, and as they are sampled more deeply, it is possible to better detect the really excellent ones. A sharp transition occurs even in this condition, exactly when the critical capacity equals 3 (brown line). As expected in this environment, the decay of the ratio between the optimal number of sampled alternatives and capacity after this transition is (slightly) faster than that of the symmetric prior. Therefore, negative skews engender a modest bias towards depth over breadth.

Finally, consider the opposite scenario, in which the prior distribution is concentrated at low success probability values ($\alpha = 1,\ \beta = 3$, positively skewed beta distribution), which corresponds to looking for a 'needle in a haystack' or a 'poor' environment. In this scenario, one ought to sample more alternatives less deeply to allow for the possibility of finding the rare good alternatives, and thus breadth should be emphasized over depth (Fig. 2.2d, blue line). In this scenario, the sharp transition occurs at capacities around 10 (blue line).

Despite the large variations of prior distributions, a fast transition occurs in all conditions at around a small capacity value, like in the uniform prior case. In addition, a power law behavior is observed at larger values of capacity regardless of skew, with exponents close to $1/2$ in all cases (uniform prior, exponent = 0.49; negatively skewed prior, 0.49; positively skewed prior, 0.64; s.e.m. = 0.01). These behaviors are observed over a larger range of parameters of the prior distribution (Fig. 2.3)

One interesting limit scenario arises for strongly positively skewed prior distributions, e.g., by taking β to infinity while fixing $\alpha = 1$. In this limit, the prior mean probability $\alpha/(\alpha + \beta)$ decreases to zero, and the critical capacity rises very steeply to infinity (Fig. 2.3a as one moves leftwards). Increasing the prior's skewness makes finding good options less likely, as most of the options are very likely to be very bad, akin to an extreme case of the haystack environment considered before. As expected, this makes breadth search optimal for increasingly larger values of capacity, as indicated by the increasing values of critical capacity. However, for large enough capacities a transition is still observed above which a roughly balanced mix between breadth and depth becomes optimal. More precisely, in this regime the optimal number of sampled alternatives features a power behavior with exponents close but above $1/2$, indicating a bias towards breadth (leftmost points

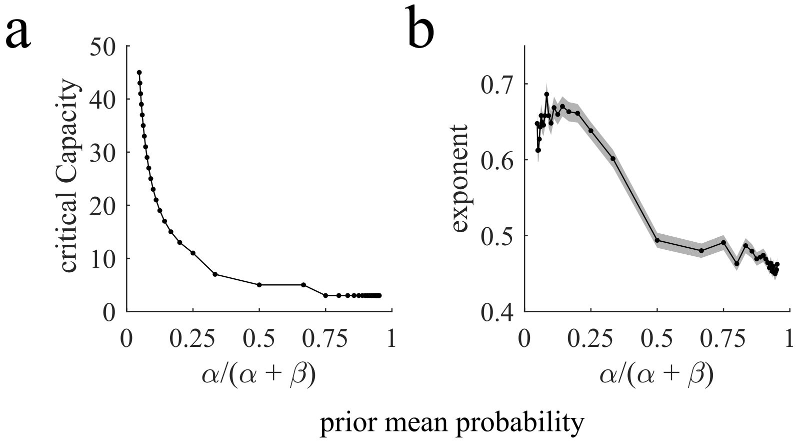


Figure 2.3: Sharp transitions at relatively low capacity values and close to square root sampling behavior for large capacity are observed for a broad range of parameters of the prior distribution. (a) Critical capacity decreases very steeply to low values (≈ 10) as a function of the prior mean probability $\alpha/(\alpha + \beta)$. (b) Exponents decrease as a function of the prior mean probability and cluster around $1/2$. The exponents are obtained from log-log linear regression fits of the optimal number of alternatives samples vs capacity ($M \propto C^{exponent}$) for C's ranging from 1000 to 2000 in steps of one. Shaded areas correspond to 95 CIs. In both panels, points are obtained by theoretical predictions from Eq. 4 in the Methods. For prior mean probabilities smaller than or equal to $1/2$ we fix $\alpha = 1$ while β varies from 1 to 20 in steps of one, and for values larger than $1/2$ we fix $\beta = 1$ while α varies from 1 to 20.

in Fig. 2.3b). When the prior mean probability exceeds values as low as 0.1, the critical capacity plateaus to low values below 10, and the exponent drops to values smaller but close to $1/2$, indicating a weak preference towards depth.

To test how robust the behaviors we explored are, we furthermore considered Gaussian rather than Bernoulli samples (Supplemental Fig. 2.7). Strikingly, for a large range of the samples' noisiness, we again observed a sharp transition occurring at low critical capacities (~ 10). Below the critical capacity, breadth search is preferred, while above it a mix between breadth and depth is optimal, characterized by a power law behavior (exponent = 0.35, 95% CI $= [0.30,\ 0.41]$). Thus, the resulting strategy was qualitatively identical, and numerically similar, to the Bernoulli samples case.

2.2.4 Optimal choice sets and sample allocations

So far, we have focused on optimal even sample allocation. Let us now consider the payoffs for decision-makers willing to consider all possible allocation strategies.

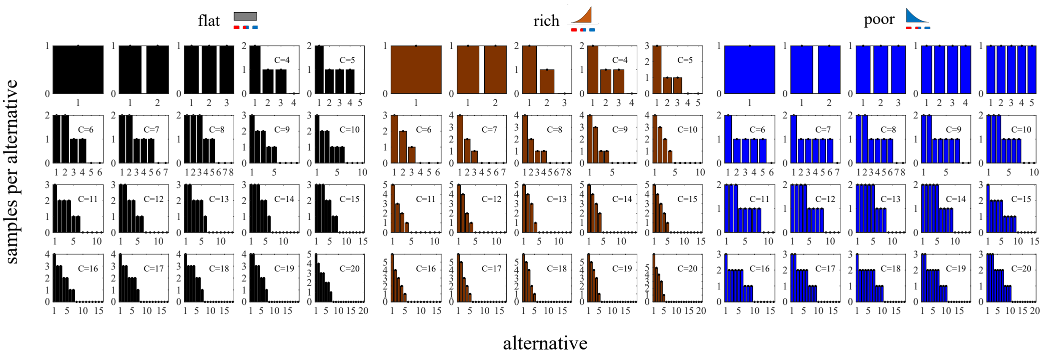


Figure 2.4: Optimal sample allocations and choice sets. Optimal sample allocation for flat, rich and poor environments from capacity $C = 1$ up to $C = 20$. The environments correspond respectively to uniform, negatively and positively skewed prior distributions (top icons). Optimal sample allocations are represented as bar plots, indicating the number of samples allocated to each alternative ordered from the most to the least sampled alternative.

The number of all possible allocations equals the number of partitions of integers in number theory, which grows exponentially with the square root of capacity (Andrews, 1998). This makes finding the globally optimal sample allocation a problem that is intractable in general. For small capacity values $C \leq 7$ and uniform prior distributions we compute the exact optimal sample allocation by exhaustive search and rely on a stochastic hill climbing method for larger capacities and other priors. The latter finds a local maximum for the utility, which is likely to be a global maximum, as we found it to coincide with the one provided by exhaustive search for small capacities $C \leq 7$, and the optimal utility did not significantly change across different initializations and random seeds for larger capacity values.

Globally optimal sample allocation (which defines optimal choice sets) for a uniform prior beta distribution tends to sample all or most of the alternatives when the capacity is small, but as capacity increases the number of sampled alternatives decrease (Fig. 2.4, left). For instance, for capacity equal to 5 samples, the optimal sample allocation is (2,1,1,1,0). In general, in optimal allocations, the decision-maker adopts a local balance between oversampling a few alternatives and sparsely sampling others –a local compromise between breadth and depth— even though all options are initially indistinguishable. This further level of specialization and distinction between alternatives might be able to better break ties between similar alternatives when compared to an even sampling strategy.

We also studied optimal sample allocation for positively and negatively skewed prior distributions. In a rich environment (center panel), the optimal sample allocation is uneven for capacities as low as $C = 3$. In contrast, in a poor environment (right), the optimal sample allocation remains even up to capacity C =5, which was not the case for the flat environment (compare with left panel). For higher capacities, decision-makers in rich environments ought to sample less broadly but more deeply. For instance, for capacity $C = 20$, only around 5 alternatives are sampled, while the remaining 15 potentially accessible alternatives are neglected. In the haystack environment, in contrast, about half of the alternatives are sampled, but not very deeply (only a maximum of 3 samples are allocated to the most sampled alternatives).

2.2.5 Even sample allocation is close to optimal

Three principles stand out. First, globally optimal sample allocation almost never coincides with optimal even allocation. Second, at low capacity optimal allocation favors breadth while at large capacity a breadth-depth balance is preferred (Fig. 2.5a). Third, a fast transition is observed between the two regimes happening at a relatively small capacity value. The last two features are shared by the optimal even allocations as well (cf. Fig. 2.2c).

Optimal even and globally optimal sample allocations share some important features, but are they equally good in terms of average reward obtained? We compared the average reward from globally optimal and even optimal sample allocations. For comparison, we always used even sampling based on a uniform prior over each alternatives' success probabilities, that is, we sample $M = C$ alternatives with one sample each if capacity is $C \leq 7$ and $M = \sqrt{C}$ alternatives with $L = \sqrt{C}$ samples each if capacity is larger (square root sampling law; see Methods for details). This heuristic produced comparable performances to the optimal ones (Fig. 2.5b). The worst-case scenario occurred in the poor environment (blue line) when capacity is close to 10, which led to a drop in reward by close to 10%, but the maximum discrepancy value was even smaller for the flat and rich environments. Indeed, for the flat environment, the maximum drop in reward was only around 5%.

For large capacity $C > 100$ the square root sampling law produced results that were very close to the performance of the optimal solutions (as found by stochastic hill climbing). Therefore, the gain of globally optimal sample allocation over optimal even sampling at low capacity, and over the square root sampling law for high capacity, is at most marginal.

We also compared the merits of the square root sampling law to other sensible heuristics: pure breadth, pure depth, random sampling of options and a triangular approximation. Pure breadth search allocates just one sample per alterative, such that the number of sampled alternatives equals capacity. The pure depth heuristic randomly chooses two alternatives that are each allocated half of the sampling capacity. Random search randomly assigns each of the C samples to any alternative with replacement. A final heuristic, called 'triangular', is inspired by the seemingly

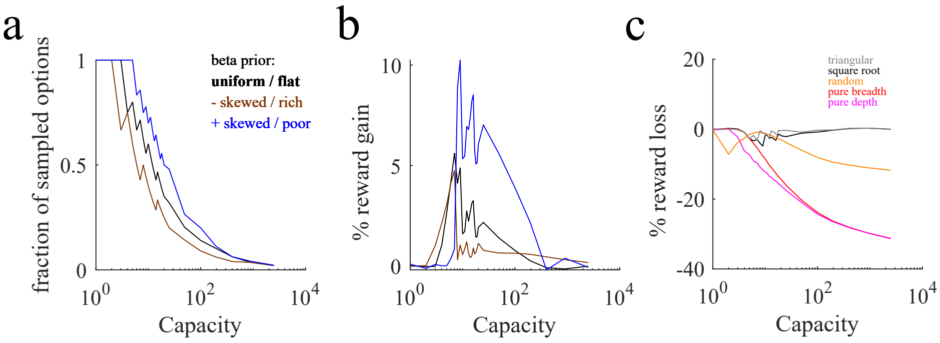


Figure 2.5: Globally optimal and optimal even sample allocations share similar features and have similar performances. (a) Fraction of sampled options (compared to the maximum number of potentially accessible alternatives, equal to C) as a function of capacity C. The fraction is close to one for small values for all environments (flat -black line, rich -brown, and poor -blue). The fraction decays rapidly to zero from a critical value that depends on the prior. The jagged nature of the lines is due to the discrete nature of capacity. (b) Percentage increase (gain) in averaged reward by using globally optimal sample allocation compared to even allocation (see Appendix). Color code as in the previous panel. (c) Percentage loss in averaged reward by using triangular (gray), square root sampling law (black), random (orange), pure breadth (red) and pure depth (pink) heuristics compared to optimal allocation, in a flat environment. For the square root and triangular heuristics, we used pure breadth search when $C \leq 5$.

isosceles right triangle shape of the optimal allocations (see Fig. 2.4). It splits capacity by sampling the first alternative with $\left\lfloor 2\sqrt{C} \right\rfloor - 1$ samples and any further alternative with one sample less than the previous one until capacity is exhausted ($\lfloor x \rfloor$ is the floor function). All heuristics finally choose the alternative with the highest posterior mean probability. While the loss relative to optimal allocation is smallest for triangular allocation, the square root sampling law performs similarly, and much better than random, pure breadth and pure depth heuristics (Fig. 2.5c).

2.2.6 Dynamic allocation of capacity

Thus far, we have considered 'static' allocations whereby no feedback is provided before all samples are allocated. In a less rigid 'dynamic' sample allocation strategy some basic form of interim feedback might be available, based upon which further alternatives can be sampled. To model such a scenario, assume that the capacity can be divided into a sequence of a maximum of C waves $k = 1, \ldots, C$, such that in each wave a number of alternatives M_k, no larger than in the previous wave, is sampled just once. The number of alternatives sampled at each wave can be chosen freely, but has to be allocated before sampling starts, that is, the decision maker has to determine the policy at the start knowing she will receive feedback in the future. However, to dynamically react to past sampling outcomes, the kth wave allocates its M_k samples to only those M_k alternatives with the largest number of successes so far (with random allocation in case of ties). This implies that in wave $k + 1$ one can only sample a subset of the alternatives sampled in wave k. Once sampling has been completed across all waves, the alternative with the highest posterior mean probability is chosen among the M_1 sampled alternatives in the first wave. We restricted the final choice to this initial set of alternatives to handle the unlikely case that the lastly sampled alternatives turned out to be worse than (our a priori belief about) the initially sampled ones. In that case, the dynamic strategy could lead to worse performance than the static one. We call the above flexible allocation of the predefined sample waves 'dynamic' sample allocations. As for static allocations, we find the best-performing $M_1, M_2, \ldots$ sequence by stochastic hill climbing.

Optimal dynamic sample allocations share many features with optimal static sample allocations (Fig. 2.6). At low capacity, pure breadth search is again optimal. That is, it is best to allocate all samples in the first wave, assigning just one sample per alternative (Fig. 2.6a). For capacities larger than the critical capacity C=3, it is best to mix breadth with depth search, and for very large capacity most accessible alternatives are again ignored. The optimal dynamic and static sample allocations have, however, important differences (Fig. 2.6b; cf. Fig. 2.4). Specifically, the initial wave tends to sample many alternatives to identify good ones, and follow-up waves narrow down the search to the potentially best ones. This results in broader sample allocations (Fig. 2.6c), that, overall, sample more alternatives than for static allocations (cf. Fig. 2.4). Finally, we test how the static square root sampling law performs against the optimal dynamic allocations, finding that the former is worse by less than 9% for all capacity values (Fig. 2.6d). We also confirm that static random, pure breadth and pure depth strategies are substantially worse than the square root sampling law, while the triangular strategy is similar to the simple square root sampling heuristic.

2.3 Discussion

We delineate a formal mathematical- framework for thinking about a commonplace decision-making problem. The *breadth-depth dilemma* occurs when a decision-maker is faced with a large set of possible alternatives, can query multiple alternatives simultaneously with arbitrary intensities, and has overall a limited search capacity. In such situations, the decision-maker will often have to balance between allocating search capacity to more (breadth) or to fewer (depth) alternatives. We

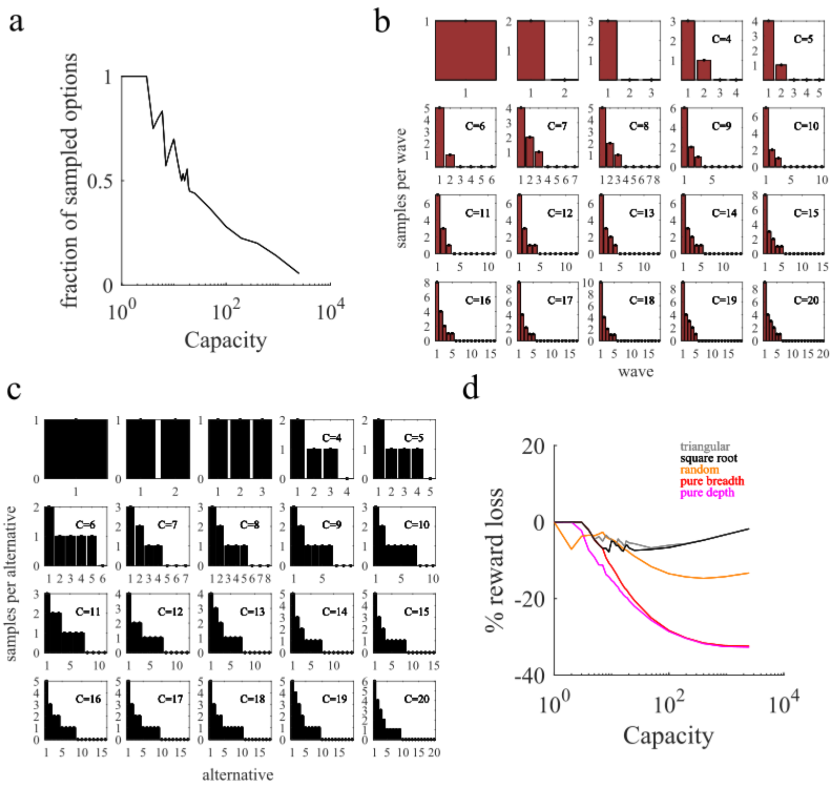


Figure 2.6: Optimal dynamic sample allocations display a sharp transition at low capacity, distribute samples unevenly across alternatives, and ignore a vast number of alternatives at high capacity. (a) Fraction of sampled alternatives (compared to the maximum number of potentially accessible alternatives, equal to C) as a function of capacity C for the flat environment (uniform prior). The fraction is one for small capacity values and decays rapidly to zero at large capacity. (b) Optimal sample waves, indicating the number of samples allocated in each wave. The number of samples allocated in each wave lies between 1 and C and they sum up to the total available capacity C. The maximum allowed number of waves is C. (c) Optimal dynamic sample allocations and choice sets after the whole capacity has been allocated through the sample waves. The alternatives with largest number of successes are allocated a higher number of samples compared to static allocations (cf. Fig. 2.4). Many alternatives are given just one sample, typically arising from the first wave, which produces broader sample allocations compared to static allocations. (d) Percentage loss in averaged reward by using triangular (gray), square root sampling law (black), random (orange), pure breadth (red) and pure depth (pink) static heuristics compared to optimal dynamic allocation.

develop and use a finite sample capacity model to analyze optimal allocation of samples as a function of capacity. The model displays a sharp transition of behavior at a critical capacity corresponding to just a small set of available samples (around 10). Below this capacity, the optimal strategy is to allocate one sample per alternative to access as many alternatives as possible (i.e. breadth is favored). Above this capacity, breadth-depth balance is emphasized, and the *square root sampling law*, a close-to-optimal heuristic, applies. That is, capacity should be split into a number of alternatives equal to the square root of the capacity. This heuristic provides average rewards that are close to those from the optimal allocation of samples. As it is easy to implement, it can become a general rule of thumb for strategic allocation of resources in multi-alternative choices. The same results roughly apply to a wide variety of environments, including flat, rich and poor ones, characterized by very different difficulties of finding good alternatives.

Despite the billions of neurons in the brain, our processing capacity seems quite limited. This strict limit applies to attention, where it is sometimes called the attentional bottleneck (Deutsch and Deutsch, 1963; Treisman, 1969; Yantis and Johnston, 1990), including spatial attention, where the limit is best characterized (Desimone and Duncan, 1995), over working memory (Brady et al., 2011; Cowan et al., 2005; Luck and Vogel, 2013; Ma et al., 2014; Miller, 1956; Sims, 2016), to executive control (Norman and Shallice, 1986; Shenhav et al., 2017; Sleezer et al., 2016) and to motor preparation (Cisek and Kalaska, 2010) These narrow limits, which often number only a handful of items (though see (Ma et al., 2014)), suggest some sort of bottleneck. However, another interpretation is that capacity is much larger than it appears, and instead, observed capacity reflects the strategic allocation of resources according to the compromises that our model identifies as optimal. The square root sampling law, in other words, suggests that the apparently narrow bandwidth of cognition may reflect the optimal allocation across very few alternatives of a relatively large capacity.

This is particularly likely to be true for economic choice. We are especially interested in the apparent strict capacity limits of prospective evaluation (Hayden and Moreno-Bote, 2018; Krajbich et al., 2010; Lim et al., 2011; Redish, 2016; Rich and Wallis, 2016). Indeed, the failures of choice with choices sets over a few items are striking and have been a major part of the heuristic literature (Diehl and Poynor, 2010; Iyengar and Lepper, 2000). These strict limits are ostensibly difficult to explain. They do not appear to derive, for example, from the basic computational or biophysical properties of the nervous system, as is evident from the fact that our visual systems are an exception to the general pattern and can process much information in parallel. Nor do these limits appear to relate to any desire to reduce the extent of computation, as large numbers of brain regions coordinate to implement these cognitive processes (Rushworth et al., 2011; Siegel et al., 2015; Vickery et al., 2011; Yoo and Hayden, 2018). Our results presented above offer an appealing explanation for this problem: economic choice can be construed as breadth-depth search problems, and even when capacity is large, the optimal strategy is to focus on a very small region of the search space. Thus, our results can also help to understand why many cognitive systems operate in a regime of low sampling size, thus resolving the paradox of why low breadth sampling and large brain resources can coexist.

We believe that these results are particularly relevant to behavioral economics. Research has shown that consumers often consider just a small number of brands from where to purchase a specific product out of the many brands that exist in the market (Hauser and Wernerfelt, 1990; Stigler, 1961). The prevailing notion is that decision-makers hold a consideration choice set from where to make a final choice rather than contemplating all possibilities. Several reasons for this behavior have been provided. First, choice overload has been shown to produce suboptimal choices in certain conditions (Iyengar and Lepper, 2000; Scheibehenne et al., 2010). Secondly, selecting a small number of options from where to choose can be actually optimal if there is uncertainty about the value of the options and there is cost for exploring and sampling further options (Mehta et al., 2003; Roberts and Lattin, 1991; Santos et al., 2012). Estimating the overall benefits of considering larger sets has to be balanced with the associated cost of exploring further options.

This research has provided a relevant line of thought for understanding low sampling behavior within the context of bounded rationality by formally assuming the presence of linear costs of time for searching for new options. Time costs come in their models at the expense of unknown parameters, which often are difficult to fit (Mehta et al., 2003; Roberts and Lattin, 1991). Further, linear time costs always permit unlimited number of sampled options, as they do not impose a strict limit in the number of options that can be sampled. In our approach, in contrast, allocating finite resources imposes a strict limit to the number of options that can be sampled and, as resources are limited, there is a tradeoff between sampling more options with less resources or sampling fewer options with more resources, directly addressing the breadth-depth dilemma. This difference could be the main reason why the consideration set literature has not reported sharp transitions of behavior as a function of model parameters (costs) nor power sampling laws, which are the main features of our finite sample capacity model.

A number of extensions would be required to fully address more realistic problems associated to the breadth-depth dilemma. So far, we have considered a two-stages decision process, where the first metareasoning decision is about optimally distributing limited sampling capacity. We have also considered a sequential problem where some basic form of feedback can be used, but the allocation strategy needs to be chosen before the gathering of information and remains fixed thereafter. By construction, these optimal dynamic allocations at large capacity sample more deeply those alternatives that have largest values, in line with experimental work

(Krajbich et al., 2010; Sepulveda et al., 2020). Perhaps a more relevant observation is that the depth of processing of the best alternatives increases with capacity and that more samples are allocated to the top alternatives than for optimal static sample allocations (cf. Fig. 2.6). Furthermore, if capacity increases, relatively more samples are allocated to the most-sampled than the second-most-samples alternative. Both predictions are currently untested. It would be interesting to extend these results to truly sequential processes where the decision of how many samples to allocate per wave is flexible and depends on intermediate feedback. An advantage of this more general setup (Morgan and Manning, 1985) is that a full-fledged interaction between the breadth-depth and exploration-exploitation dilemmas could be studied. In particular, a relevant direction is relating our square root sampling law with Hick's law (Hick, 1952) for multialternative choices. The two approaches touch different aspects of multialternative decision making: while Hick's law refers to the problem of how long options should be sampled in a multialternative setting, it does so by sampling all available options; the square root sampling law, by contrast, applies to situations where there are many alternatives and a large fraction of them are to be ignored due to limited capacity, directly facing the breadth-depth dilemma. It will be interesting to integrate the two sets of results within a general framework of multialternative sequential sampling (Roe et al., 2001; Tajima et al., 2016; Usher and McClelland, 2004) under limited resources.

A second possible extension of our work is reconsidering the nature of capacity. For instance, 'rate distortion theory' defines a natural capacity constraint over the mutual information between the inputs and the outputs in a system (Bates et al., 2019; Sims, 2003). This capacity constraint might more naturally enforce a finite capacity than fixing the total number of samples that a system can draw from (externally or internally). A third relevant direction would be extending our study to cases where the capacity is continuous rather than discrete, and to cases where the observations are continuous variables. Showing that Gaussian rather than Bernoulli outcomes yield qualitatively similar strategies is a first step in this direction. Although it remains a topic for future research, we do not expect qualitative differences in behavior in other continuous settings, as for large capacity the continuous limit approximation applies, and for low capacities the optimality of low number of alternatives is expected.

While we do not know of direct tests of breadth-depth capacities in humans, indirect measurements suggest that the square root sampling law can be at work in some realistic conditions, such as chess. It has been argued that chess players can image around 100 moves before deciding their next move (Simon, 1972). Assuming that their capacity is 100, then the square root sampling law would predict that players should sample 10 immediate moves followed by around 10 continuations. Indeed, estimates indicate that chess players mentally contemplate roughly between 6-12 immediate moves followed by their continuations (Simon,

1972) before capacity is exhausted due to time pressure. Although decisions in trees like this surely involve other types of search heuristics beyond balancing breadth and depth, the quantitative similarity between predictions and observations is intriguing.

Finally, our work potentially opens ways to understand confirmation biases. Confirmation biases happen when people extensively sample too few alternatives, thus effectively seeking information for the same source. We have demonstrated that oversampling some alternatives and completely ignoring others is optimal in certain conditions. It remains to be seen, however, whether this is actually the optimal strategy under more general conditions or whether the oversampling strategy induces severely harmful biases in certain niches.

It is important to note that we have described the phenomenology of the breadth-depth dilemma in conditions where all alternatives are, a priori, equally good. Thus, ignoring a large fraction of options and the associated square root sampling law can only be the worst-case scenario, in the sense that if there are biases or knowledge that a subset of alternatives is initially better than the rest, then fewer number of alternatives should be sampled. This consideration reassures us in the conclusion that the number of alternatives that ought to be sampled is much smaller than sampling capacity, an observation that might turn to be of general validity in both decision-making setups as well as in terms of brain organization for cognition.

2.4 Appendix

2.4.1 Finite capacity model

We consider a two-stage decision process in a multi-alternative decision-making problem modeled as a partially observable Markov decision process. There are N alternatives, defined each by a Bernoulli random processes, whose trial by trial ($t = 1, ...$) outcomes follow $s_i^t \sim \mathrm{Bern}\,(p_i)$, $s_i^t \in \{0, 1\} = \{\mathrm{failure}, \mathrm{success}\}$, $i = 1, ..., N$. The outcomes are independently distributed for all trials t and across alternatives. The values of the success probabilities are unknown to the decision-maker, and follow a prior distribution $p_i \sim \mathrm{Beta}(\alpha, \beta)$ i.i.d. for all alternatives, with known hyperparameters (α, β). Allowed actions follow a two-stage decision process. In the first stage, the decision-maker can draw a total of $C = N$ samples at once, namely, a one-go decision is considered (Bechhofer and Kulkarni, 1984; Gupta and Liang, 1989; Sobel and Huyett, 1957). We consider the case where the total number of alternatives N exhausts sampling capacity C, but the results are equivalent if the number of alternatives is larger than capacity, with the addition of more rejected or non-sampled alternatives. The action space is $\mathsf{A}_\mathsf{L} = \{\vec{L} : L_i \geq 0 \ \forall i, \sum_i L_i = C\}$, where $\vec{L} = (L_1, ..., L_N)$ is the number of samples drawn from each of the alternatives, with the constraint $\sum_i L_i = C$ (we often refer to the vector $\vec{L}$ as sample allocation). Note that the decision-maker can decide to sample the same alternative several times (i.e., $L_i > 1$ for some i), and also decide not to sample from several alternatives (i.e., $L_i = 0$ for other i). In general, $M \leq C = N$ alternatives are sampled. If just a few alternatives are sampled ($M \sim 1$), many samples can be allocated to each. If C alternatives are sampled, only one sample could be allocated to each of them. Outcomes of the samples from the sampled alternatives are revealed all at once, not sequentially. In the second stage of the decision-making process, after outcomes are observed, the decision-maker should decide what alternative to choose. We initially assume that it is possible to choose only among the sampled alternatives. Thus, the action space in the second stage is defined by the set $\mathsf{A}_\mathsf{C} = \{c : L_c > 0\}$ of size M, ordered as $\{c_1, ..., c_M\}$. The sufficient statistics of the outcomes of the Bernoulli processes to infer the success probabilities are the counts of successes for each of the M sampled alternatives, $\vec{n} = (n_{c_1}, ..., n_{c_M})$, with $n_j = \sum_{t=1}^{L_{c_j}} s_{c_j}^t$, and thus the decision of what option to choose should be a function of those counts and on the sample allocation vector $\vec{L}$, which together constitute the information state of the decision process. The counts, conditioned on the success probabilities, follow $n_i \sim \mathrm{Bin}(p_i, L_i)$. Note that the dimension of the vector $\vec{n}$ depends on the number of sampled alternatives (those satisfying $L_i > 0$) and thus the consideration set changes size depending on the first stage decision.

We define the utility of a choice $i \in \mathsf{A}_\mathsf{C}$ as the hidden value of the success probability of the corresponding Bernoulli variable, $U_i = p_i$. We assume that the decision-maker maximizes expected utility. This involves determining the optimal allocation of samples $\vec{L}^*$ to be used in the first stage followed by defining an optimal decision rule that selects one of the sampled alternatives based on $\vec{n}$. A decision rules maps an observation $\vec{n}$, given the allocation vector $\vec{L}$, into an element of the action space A_C. By considering all possible decision rules, $\delta = \{\delta(\vec{n}, \vec{L}) : (\vec{n}, \vec{L}) \to \mathsf{A}_\mathsf{C}\}$, we show in Sec. (2.4.4) that the optimal decision rule, $\delta^*(\vec{n}, \vec{L})$, is the one that selects always, for any sample allocation $\vec{L}$, the alternative with the maximum posterior mean success probability $\mathbb{E}(p_i|n_i, L_i) = \frac{n_i+\alpha}{L_i+\alpha+\beta}$, $i \in A_C$, or chooses any of the maximum ones if there are ties. Therefore, the expected utility for a given sample allocation $\vec{L}$ following the optimal decision rule is

$$U(\vec{L}) = \sum_{\vec{n}} p\left(\vec{n}|\vec{L}, \alpha, \beta\right) \max_{i \in \mathsf{A}_\mathsf{C}} \left(\frac{n_i + \alpha}{L_i + \alpha + \beta}\right), \tag{2.6}$$

where the joint posterior over $\vec{n}$ factorizes into beta-binomial distributions as $p\left(\vec{n}|\vec{L}, \alpha, \beta\right) = \prod_{i \in \mathsf{A}_\mathsf{C}} \mathrm{Bb}(n_i|L_i, \alpha, \beta)$. Then, the optimal sample allocation $\vec{L}^*$ equals

$$\vec{L}^* = \arg\max_{\vec{L} \in \mathsf{A_L}} U(\vec{L}) = \max_{\vec{L} \in \mathsf{A_L}} \sum_{\vec{n}} p\left(\vec{n}|\vec{L}, \alpha, \beta\right) \max_{i \in \mathsf{A_C}} \left(\frac{n_i + \alpha}{L_i + \alpha + \beta}\right), \qquad (2.7)$$

and the corresponding maximum expected utility becomes

$$U^* = \max_{\vec{L} \in \mathsf{A_L}} U(\vec{L}). \qquad (2.8)$$

Finding the optimal solution in Eq. (2.7) is hard because of the large number of sample allocations that it is possible to form out of C samples. The number of unique partitions of C samples equals the number of integer partitions of C (not to be confused with the Bell number), for which we are not aware of simple exact expressions. We should only consider unique partitions because all the alternatives are initially (before sampling) indistinguishable. Therefore, without loss of generality, we can always assume that we sample the alternatives by using the sample allocation $\vec{L} \in \mathsf{A_L}$ where we impose the additional constraint that $L_i \geq L_{i+1}$ for $i = 1, ..., N-1$. That is, we sample the first alternative with more or the same number of samples as the second alternative, the second alternative with more or the same number of samples as the third one, and so forth. We describe a stochastic hill climbing algorithm bellow in Sec. (2.4.4) to find the optimal sample allocation exactly for small capacity C and approximately for large capacity. To find useful analytical expressions for Eqs. (2.7, 2.8), we restrict ourselves further by first looking for optimal even sample allocations, that is, allocation of samples across $M \leq C$ options with the same number of samples L per alternative. Optimal even sample allocation across alternatives is discussed in Sec. (2.4.2).

2.4.2 Analytical expressions for optimal even sample allocation

Because the space of actions $\mathsf{A_L} = \{\vec{L} : L_i \geq 0 \ \forall i, \sum_i L_i = C\}$ is very large, we restrict ourselves to a subset of possible actions, consisting in dividing the capacity C into M alternatives equally sampled with L samples each. Without loss of generality, we assume that we sample the first M alternatives and we ignore the rest of $C - M$ alternatives. Even splitting of the capacity is only possible if $C = M \times L$ holds exactly, so we will only examine the pairs (M, L) that satisfy that condition. The advantage of working in this subset of actions is that it is possible to obtain useful, exact analytical expression that will reveal non-trivial properties of the decision process. Methods for finding globally optimal sample allocation strategies are provided in Sec. (2.4.4). In the main results we also show that optimal sample allocations are not greatly better than the optimal even ones, so that even sample allocation is close-to-optimal. For an even capacity split, the optimal L^* under the constraint $C = ML$ can be obtained by specializing Eq. (2.7) to this case as

$$L^* = \arg\max_L \sum_{\vec{n}} \prod_{j=1}^{M} p\left(n_j|L,\alpha,\beta\right) \max_i \left(\frac{n_i+\alpha}{L+\alpha+\beta}\right), \qquad (2.9)$$

where $i \in \{1,...,M\}$ and $p\left(n_j|L,\alpha,\beta\right) = \mathrm{Bb}(n_j|L,\alpha,\beta)$. Naturally, the optimal number of alternatives to be sampled is $M^* = C/L^*$

A particularly simple expression results from the case $\alpha = \beta = 1$, corresponding to a uniform prior over the success probabilities of the Bernoulli variables. This is because $p\left(n_j|L,1,1\right) = \mathrm{Bb}(n_j|L,1,1) = \frac{1}{L+1}$, thus becoming a discrete uniform distribution over $n_j \in \{0,...,L\}$, independent of n_j. Then, replacing this expression in Eq. (2.9), the optimal even sample allocation simplifies to

$$\begin{aligned}
L^* &= \arg\max_L U(L), \\
U(L) &= \frac{1}{(L+1)^M} \sum_{n_1,...,n_M=0}^{L} \max_i \left(\frac{n_i+1}{L+2}\right) && (2.10)\\
&= \frac{1}{(L+1)^M(L+2)}\left((L+1)^M + \sum_{n_1,...,n_M=0}^{L} \max(n_1,...,n_M)\right) \\
&= \frac{1}{(L+1)^M(L+2)}\left((L+1)^M + \sum_{s=0}^{L} \left((s+1)^M - s^M\right)s\right) && (2.11)\\
&= 1 - \frac{\sum_{s=0}^{L}(s+1)^M}{(L+1)^M(L+2)}\,, && (2.12)
\end{aligned}$$

with $M = C/L$. Eq. (2.11) in the derivation results from realizing that the sum over $\max_i(n_i)$ contains exactly $1^M - 0$ zeros, $2^M - 1$ ones, $3^M - 2^M$ twos, etc. The sum in Eq. (2.12) is the sum of the $M - th$ powers of the first $L+1$ integers, and it can be computed using Faulhaber's formula. Eq. (2.12) confirms the intuition that the expected utility $U(L)$ for any L is smaller than one. Finally, the optimal number of evenly allocated samples (over the sampled options) can be written as

$$L^* = \arg\min_L \frac{\sum_{s=0}^{L}(s+1)^M}{(L+1)^M(L+2)} \qquad (2.13)$$

It is interesting to examine some limits of Eq. (2.12) by relaxing the constraint $C = M \times L$. For large M and $L = 1$, the expected utility in Eq. (2.12) becomes $\lim_{M\to\infty} U \to \frac{2}{3}$. This observation is not surprising, as when a very large number of alternatives is sampled with just one sample, it is very likely that at least one of them will have a successful outcome. Therefore, the expected utility of that alternative under the uniform prior will be $\frac{2}{3}$. This limit is visible in the rightmost point of Fig. 2a. In the opposite scenario, when only one alternative is sampled, $M = 1$, then the expected utility is $\frac{1}{2}$ for all L. That is, if just one alternative is sampled, then the expected probability of success of the sampled alternative is $\frac{1}{2}$, which equals the prior mean. This limit is visible in the leftmost point of Fig. 2a.

A more general way of performing the integrals involved in Eq. (2.9) is by using cumulative distribution function of the beta-binomial distributions, $F(n|L, \alpha, \beta) = \sum_{m \leq n} \mathrm{Bb}(m|L, \alpha, \beta)$. By noting that the extreme value distribution has probability mass function $F^M(n) - F^M(n-1)$ (where M denotes exponent and we have dropped conditioning to avoid cluttered notation), we can write the optimal even sample allocation in Eq. (2.9) as

$$L^* = \arg\max_L \sum_{n=0}^{L} \left[F^M(n|L, \alpha, \beta) - F^M(n-1|L, \alpha, \beta)\right] \left(\frac{n+\alpha}{L+\alpha+\beta}\right), \tag{2.14}$$

Note that the extreme value distribution $F^M(n_{max}) - F^M(n_{max}-1)$ is the distribution of $n_{max} = \max(n_1, ..., n_M)$ where $\vec{n}$ follows the above factorized beta-binomial distribution. In other words, the extreme value distribution for n_{max} is the probability that no alternative has more than n_{max} successful samples (hence the first term $F^M(n_{max})$) but removing the cases where there is no alternative with more than $n_{max} - 1$ successful samples (hence the second negative term $F^M(n_{max} - 1)$). For the uniform prior case, $\alpha = \beta = 1$, we recover Eq. (2.13), for which the cumulative can be exactly computed. For arbitrary values of α and β, Eq. (2.14) is solved numerically. These solutions are used in Fig. 2d.

The general Eq. (2.7) valid for any allocation of samples, and the specific Eq. (2.14) valid for even sample allocations, assume that a choice is made from the sampled alternatives, while non-sampled alternatives are excluded from the choice set. However, if none of the sampled alternatives turns to be good ones (e.g., because $n_i \ll L_i$ for $i \in \mathsf{A_C}$), then it would be better to choose randomly from any of the non-sampled alternatives. This is particularly so if the expected utility of any of the sampled alternatives, $\frac{n_i+\alpha}{L_i+\alpha+\beta}$, is smaller than $\frac{\alpha}{\alpha+\beta}$, which is the default expected utility of the non-sampled alternatives given that the success probabilities are drawn from a $\mathrm{B}(\alpha, \beta)$. It is straightforward to generalize these results by adding a default alternative, assumed to have utility p_0. In this case, the optimal even allocation of samples obeys

$$L^* = \arg\max_L \sum_{n=0}^{L} \left(F^M\left(n|L,\alpha,\beta\right) - F^M\left(n-1|L,\alpha,\beta\right)\right) \max\left(\frac{n+\alpha}{L+\alpha+\beta},\ p_0\right). \tag{2.15}$$

2.4.3 Asymptotic behavior for large capacity: the square root sampling law

It is possible to derive an approximation for the limiting behavior of the optimal number of sampled alternatives M^* and their associated optimal number of samples per alternative L^* by using Eq. (2.10) for large capacity C in the case of the uniform prior distribution. For large capacity C, we assume that L^* grows to infinity. This assumption is confirmed later, when the asymptotic optimal L^* is derived. If L is large, then Eq. (2.10) can be approximated by

$$\begin{aligned} U(L) &= \frac{1}{(L+1)^M} \sum_{n_1,\ldots,n_M=0}^{L} \max_i \left(\frac{n_i+1}{L+2}\right) \qquad (2.16)\\ &= \frac{1}{(L+2)}\left(1+\frac{1}{(L+1)^M}\sum_{n_1,\ldots,n_M=0}^{L}\max(n_1,...,n_M)\right)\\ &\approx \frac{1}{(L+2)}\left(1+L\int_0^1 dx_1...\int_0^1 dx_M \max(x_1,...,x_M)\right), \end{aligned}$$

where the sum in the second equation has been approximated in the third equation by an integral in the interval $[0,1]^M$ over a uniform distribution by using the transformation $n_i = Lx_i$ for $i = 1, ..., M$. The continuous approximation is valid when L is large, as assumed, since then the transformation delivers values of x_i that are dense in the unit interval. The integral can be rewritten as

$$\int_0^1 dx_1...\int_0^1 dx_M \max(x_1,...,x_M) = \int_0^1 dx_{max} x_{max} f(x_{max}),$$

where we have defined the extreme value $x_{max} = \max(x_1, ..., x_M)$. The extreme value follows the extreme value distribution $f(x_{max}) = (F(x_{max})^M)' = Mx_{max}^{M-1}$, where we have used that $F(x) = x$ is the cumulative of the continuous uniform distribution in $[0,1]$. Therefore,

$$\begin{aligned} U(L) &\approx \frac{1}{(L+2)}\left(1 + L\int_0^1 dx_{max} M x_{max}^M\right) \\ &= \frac{1}{(L+2)}\left(1 + \frac{ML}{M+1}\right). \end{aligned} \tag{2.17}$$

Finally, by maximizing $U(L)$ as a function of L with the constraint $C = ML$ we obtain the asymptotic optimal number of sampled alternatives M^* and optimal number of samples per sampled alternative L^*

$$\lim_{C\to\infty} M^* = \sqrt{C}, \quad \lim_{C\to\infty} L^* = \sqrt{C},$$

which corresponds to the square root sampling law.

In the above derivation we have assumed that L^* grows with C. To see that this corresponds to the only valid assumption to obtain L^*, let us assume now that L^* does not grow with C, that is, it is a constant or decreases with C. For any fixed value L, using Eq. (2.12) we see that $U(L) \leq 1 - 1/(L+2)$. This utility is smaller than the one obtained by using the square root law, which converges to 1, as can be easily derived from Eq. (2.17). Therefore, the square root law delivers the highest utility.

2.4.4 Optimal sample allocation

For low capacity $C \leq 7$ we found the globally optimal sample allocation strategy by exhaustive search over all possible sample allocations. For larger capacity, we searched the optimal sample allocation by using stochastic hill climbing. With this method, we confirmed that for values up to $C \leq 20$ the globally optimal sample allocations were correct up to a precision in expected utility of 10^{-4}.

We started the algorithm by using even sample allocation using the square root law heuristic: if $C \leq 7$ all options were sampled with one sample, and if capacity was larger we used the square root law by sampling $\sqrt{C}$ alternatives $\sqrt{C}$ times each. We considered the possibility that the resulting square root was not an integer, and thus we allocated the residual number of samples to a randomly chosen additional alternative; we call this allocation scheme 'even allocation'. At every iteration, we computed the expected utility of the current best sample allocation $\vec{L}$ through a Monte Carlo simulation of the Bernoulli variables and averaging utility over 4×10^5 repetitions for $C \leq 20$ and 5×10^4 for larger capacity values. A perturbed sample allocation was proposed by randomly selecting two alternatives. One sample was removed from the first alternative and added to the second one, but only if the first alternative had already assigned at least one sample. To exploit symmetry, we only consider changes of one sample from one alternative i to another $j > i$ if $L_{j-1} \geq L_j$ and $L_i \geq L_j$. If $j < i$, there were not restrictions.

With the proposed perturbed sample allocation, we computed the expected utility using the same Monte Carlo method. If the new expected value was larger than the previous one, then the proposed perturbed sample allocation became the current best sample allocation. This process was iterated 2×10^4 times for $C \leq 20$ and 3×10^3 for larger capacity values. Because at each iteration we reevaluate the expected value of the current best sample allocation, we avoid the possibility of getting stuck in a random fluctuation leading to a spuriously large expected value. The stochastic hill climbing method found optimal sample allocations that were identical to those found with the exhaustive search for low capacity $C \leq 7$. Although we do not know whether the found optimal sample allocation corresponds to a global maximum when capacity is larger, we confirmed that the optimal sample allocations found were stable against different random number seeds and initial conditions. Figs. 4 and 5 use the above method. Percentage reward gain in Fig. 5b is computed as $100 \times (U^* - U_{even})/U_{even}$, where U^* is the utility estimate of the globally optimal allocation and U_{even} is the estimate of the initial even allocation. Percentage reward loss in Fig. 5c is computed as $100 \times (U_{heuristic} - U^*)/U^*$, where U^* is the utility estimate of the globally optimal allocation and $U_{heuristic}$ is the utility estimate from triangular, square root sampling law, pure breadth or pure depth heuristics.

We also employed another version of stochastic hill climbing that avoided using extensive sampling of the Bernoulli variables to estimate expected utility. This method was used to confirm robustness of the previous results. We define the optimal utility as

$$U^* = \max_{\vec{L}} \sum_{\vec{n}} p(\vec{n}|\vec{L}, \alpha, \beta) \max_i \left(\frac{n_i + \alpha}{L_i + \alpha + \beta} \right). \tag{2.18}$$

We thus can design a Markov Chain Monte Carlo method to sample from the probability distribution

$$p(\vec{n}|\vec{L}, \alpha, \beta) = \prod_j Bb(n_j|\vec{L}, \alpha, \beta)$$

appearing in the sum of Eq. (2.18) as follows (these samples can be then used to approximate the sum). Detailed balance imposes that the probability of transitioning from a state with $\vec{n}$ to $\vec{n}'$ is the same as the converse,

$$P_{\vec{n},\vec{n}'}\, p(\vec{n}|\vec{L}, \alpha, \beta) = P_{\vec{n}',\vec{n}}\, p(\vec{n}'|\vec{L}, \alpha, \beta).$$

By proposing a change to a single alternative $n'_j = n_j \pm 1$, we can get a simple expression for the acceptance rate $r(\vec{n} \rightarrow \vec{n}')$. If $n'_j = n_j + 1$ the acceptance rate is

$$r(\vec{n} \rightarrow \vec{n}') = \min\left(1, \frac{(n_j + \alpha)(L_j - n_j)}{(L_j - n_j + \beta + 1)(n_j + 1)}\right),$$

while if $n'_j = n_j - 1$, it becomes

$$r(\vec{n} \rightarrow \vec{n}') = \min\left(1, \frac{(L_j - n_j + \beta)n_j}{(n_j + \alpha - 1)(L_j - n_j)}\right),$$

where we have made use of the Metropolis-Hastings algorithm. These two changes are proposed with equal probability and randomly across all the options. Utilities are estimated using 10^6 samples. The search over $\vec{L}$ is made using $50 \times C$ iterations. Results in Fig. 4 were reproduced by this method.

For the optimal dynamic allocations described in Fig. 6, we employed again a stochastic hill climbing method identical to the one described at the start of this section by using the vector of numbers of allocated samples per wave, M_i, $i = 1, 2, ...C$, instead of the number of samples per alternative, L_i. The method proceeded by proposing a new vector of waves $\vec{M}$ by adding a sample to a randomly chosen wave and removing a sample from another randomly chosen wave. This was done only if the second wave had at least one allocated sample to it and if the resulting proposed perturbed allocation satisfied the constraint $M_{i+1} \leq M_i$ for all i. The number of iterations and samples for Monte Carlo utility estimates are the same as above. Optimal dynamic allocations found are correct up to a precision of 10^{-4} in the utility estimates. Very similar results to those described in Fig. 6 are found when options to be sampled in each wave are selected based on their current posterior mean probabilities instead of their current number of total successes. Percentage reward loss of static heuristics compared to optimal dynamic allocations described in Fig. 6d are computed as in Fig. 5c.

Consistency

Perhaps intuitively, but wrongly, we might assume that by always opting for the alternative with larger number of successful outcomes (larger n_i in Eq. (2.7)), this would result in 'cherry picking', that is, in selecting a spuriously good option. This, in turn, would mean that we would obtain a reward that is lower than the expected utility in Eq. (2.8). Here we show, however, that the decision rule of choosing always the alternative with the highest posterior mean is both optimal and delivers on average a reward that is equal to the expected utility. This is a well-known result in statistical decision theory (Moreno-Bote, 2010; Beck et al., 2008; Berger and Berger, 1985). Here we show the derivation for completeness.

Consider any possible decision rule $\vec{d} = \delta(\vec{n})$ that assigns the counts of successes for the M sampled alternatives, $\vec{n}$, to a decision $\vec{d} \equiv \vec{d}(\vec{n}) = (d_{c_1}(\vec{n}), ..., d_{c_M}(\vec{n}))$, encoded as a one-hot vector of length M (i.e., $d_{c_i} = 1$ if alternative c_i is chosen, and $d_{c_i} = 0$ otherwise; we omit the potential dependence of the decision rule on $\vec{L}$ to avoid cluttered notation). If the success probabilities of the sampled alternatives, $\vec{p}$, are known, then by using the decision rule δ the decision-maker would have an expected utility

$$U(\vec{p}, \vec{L}, \delta) = \sum_{\vec{n}} \prod_{i \in \mathsf{A_C}} \mathrm{Bin}(n_i | L_i, p_i)\ p_i^{d_i(\vec{n})},$$

where $\vec{L}$ is the allocated number of samples over the alternatives. Note that the expected utility is an average over the values of the chosen p_i given the decision rule averaged across all possible outcomes given the allocated number of samples over alternatives. As probabilities are unknown, they are marginalized out with their prior beta distributions, resulting in the overall expected utility

$$U(\vec{L}, \delta) = \sum_{\vec{n}} \prod_{i \in \mathsf{A_C}} \frac{\Gamma(L_i + 1)}{\Gamma(n_i + 1)\Gamma(L_i - n_i + 1)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(n_i + \alpha + d_i)\Gamma(L_i - n_i + \beta)}{\Gamma(L_i + \alpha + \beta + d_i)}. \tag{2.19}$$

We note that for each term in the sum over $\vec{n}$, there is only one value of i for which $d_i = 1$ in the product, while $d_j = 0$ for $j \neq i$. The term i in the product with $d_i = 1$ gives an extra factor $\frac{n_i + \alpha}{L_i + \alpha + \beta}$ (by expanding the gamma functions just one step) that is not present in the product terms with $d_j = 0$. Therefore, the product is maximized iff $d_i = 1$ for the alternative i with maximum $\frac{n_i + \alpha}{L_i + \alpha + \beta}$ (if the maximum is not unique, any alternative with the maximum value will give exactly the same result). This result proves that the optimal decision rule δ^* is the one that chooses always the alternative with the highest posterior expected utility given $\vec{n}$.

Now, we can show that for the optimal decision rule δ^*, the expected utility is the same as that in Eq. (2.8). We can rewrite Eq. (2.19) as

$$U(\vec{L}, \delta^*) = \sum_{\vec{n}} \max_{i \in \mathsf{A_C}} \left(\frac{n_i + \alpha}{L + \alpha + \beta} \right) \prod_{i \in \mathsf{A_C}} \frac{\Gamma(L_i + 1)}{\Gamma(n_i + 1)\Gamma(L_i - n_i + 1)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(n_i + \alpha)\Gamma(L_i - n_i + \beta)}{\Gamma(L_i + \alpha + \beta)},$$

which is identical to the maximum expected utility $U(\vec{L})$ in Eq. (2.8), that is, $U(\vec{L}, \delta^*) = U(\vec{L})$. This shows that 'cherry picking' is optimal.

2.4.5 Supplementary Figures

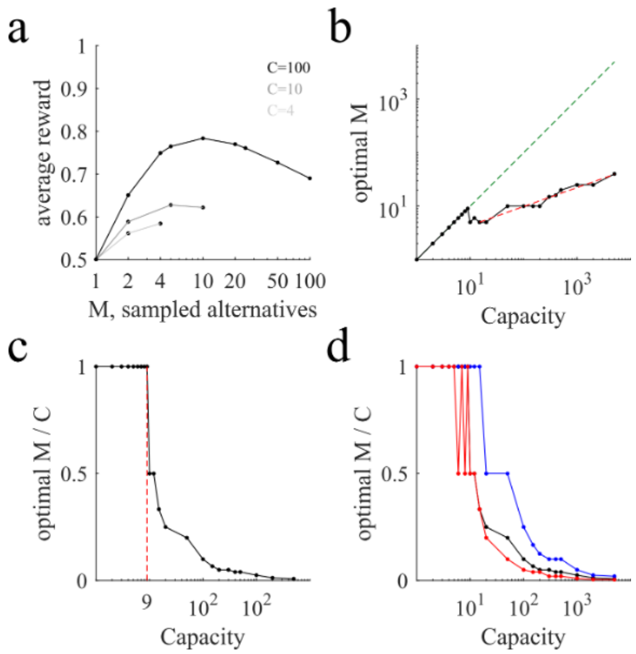


Figure 2.7: Sharp transitions in optimal number of sampled alternatives at low capacity and power law behavior at high capacity in a breadth-depth (BD) model with Gaussian outcomes. Each option is modelled as a Gaussian with known variance σ^2 and unknown mean reward μ drawn independently for each option from a uniform prior over $[0, 1]$. The goal is to optimize the even allocation of C independent samples over at most C options to maximize the posterior mean reward of the best option. The less options samples are allocated to, the better the estimates of the underlying means of the sampled options. The optimal even allocations observed qualitatively match those found in Fig. 2.1 for Bernoulli observations with unknown success probabilities. (a) Average reward (points and lines, simulations) as a function of the number of sampled alternatives M for three different capacities ($C = 4, 10, 100$; light, intermediate and dark lines respectively) for variance $\sigma^2 = 1$, comparable to prior's width. The maximum occurs at the right extreme for low capacity but at a relatively low values for large capacities. Note log horizontal scale. (b) Optimal number of sampled alternatives as a function of capacity. When capacity is smaller than around 9, a linear trend of unit slope is observed (dashed green line), but when capacity is above 9, the behavior becomes sublinear (dashed red line corresponds to the best power law fit, with exponent close to $1/3$; power law fit, exponent = 0.35, 95% CI = $[0.30, 0.41]$). The transition between these two regimes is sharp. (c) The sharp transition is clearer when plotting the optimal number of sampled alternatives to capacity ratio as a function of capacity. For low capacity, the ratio is one, but for large capacity the ratio decreases very rapidly. The last point below which the optimal ratio is always one (critical capacity) corresponds to capacity equal to 9 (indicated by the vertical red line). (d) Number of sampled alternatives to capacity ratios for different variances $\sigma^2 = 0.1, 1, 10$ (blue, black, red lines, respectively), corresponding to reliable, standard, and unreliable Gaussian samples. All points and lines correspond to simulations. When samples are reliable (blue line), breadth search is favored, as can be seen from the increase of the critical capacity and the slower decay of the optimal ratio M/C. In contrast, when samples are unreliable (red line), depth search is favored.

Chapter 3

BREADTH-DEPTH DILEMMA FOR CONTINUOUS RESOURCES

The following chapter is based on the manuscript published in the journal of Cognitive Science, see (Ramírez-Ruiz and Moreno-Bote, 2022) for bibliographic details. I declare to be the main author of this work.

Abstract

When facing many options, we narrow down our focus to very few of them. Although behaviors like this can be a sign of heuristics, they can actually be optimal under limited cognitive resources. Here we study the problem of how to optimally allocate limited sampling time to multiple options, modelled as accumulators of noisy evidence, to determine the most profitable one. We show that the effective sampling capacity of an agent increases with both available time and the discriminability of the options, and optimal policies undergo a sharp transition as a function of it. For small capacity, it is best to allocate time evenly to exactly five options and to ignore all the others, regardless of the prior distribution of rewards. For large capacities, the optimal number of sampled accumulators grows sub-linearly, closely following a power law as a function of capacity for a wide variety of priors. We find that allocating equal times to the sampled accumulators is better than using uneven time allocations. Our work highlights that multi-alternative decisions are endowed with breadth-depth tradeoffs, demonstrates how their optimal solutions depend on the amount of limited resources and the variability of the environment, and shows that narrowing down to a handful of options is always optimal for small capacities.

3.1 Introduction

The problem of allocating finite resources to determine the best of several options is common in decision making, from deciding which vaccine candidates to fund for further research to choosing a movie for Saturday night. In these cases, planning, and thus resource allocation, needs to be made in advance, well before feedback about the success of the choice is observed. Consequently, two important questions arise: How many options should we examine? And, for how long? When resources are limited, such as number of participants that can be tested with vaccines in a short time, or weekend free time in the previous examples, a decision maker should balance breadth, how many options to sample, and depth, how much to sample each. This ubiquitous decision making problem under constrained resources is what has been called the breadth-depth dilemma (Miller, 1981; Horowitz and Sahni, 1978; Moreno-Bote et al., 2020).

In the face of many alternatives, humans quickly narrow down the number of considered options to around two to five (Payne, 1976; Olshavsky, 1979; Beach, 1993; Levin et al., 1998; Hauser and Wernerfelt, 1990), and, when presented with more than six options, experienced overload produces suboptimal choices in certain conditions (Iyengar and Lepper, 2000; Scheibehenne et al., 2010). Models describe this behavior by assuming that considering more options incurs search or mental costs (Hauser and Wernerfelt, 1990; Mehta et al., 2003; Stigler, 1961), but why people consider small sets in a wide range of environments is still a matter of debate. While this could be explained by strict small capacity limits in attention or working memory (Miller, 1956; Cowan et al., 2005), the nature of this small capacity would still need to be addressed (Ma et al., 2014). Another possibility is that capacity is not necessarily small, but rather that sampling few options and ignoring the vast majority, in either an automatic or in a conscious manner, is actually an optimal policy that favors depth over breadth (Moreno-Bote et al., 2020). This possibility is supported by the fact that neuronal resources devoted to decision making are not precisely low, as dozens of brain areas and several billions of neurons are involved in even simple decision making tasks (Rushworth et al., 2011; Siegel et al., 2015; Vickery et al., 2011; Yoo and Hayden, 2018). Thus, processing bottlenecks could be reflections of close-to-optimal policies to breadth-depth dilemmas.

Bounded rationality accounts (Simon, 1972; Russell and Wefald, 1991; Gershman et al., 2015; Griffiths et al., 2015) surmise that many features of cognition arise from the finite limits of the nervous system. This must also be the case for the nature of the policies chosen by people in decision making, but oftentimes the constraints imposed by the limited resources are not made explicit. Indeed, choices stemming from sequential sampling between two or three options have been typically modelled as optimal stopping problems (Ratcliff and Murdock, 1976; Gold and Shadlen, 2007; Krajbich and Rangel, 2011; Drugowitsch et al., 2012; Tajima et al., 2019; Jang et al., 2021; Callaway et al., 2021; Vul et al., 2014; Sepulveda et al., 2020), where agents should optimally balance the prospect of learning the value of the options with the costs of sampling them, but they do so without computational or capacity constraints. In these works, the objective is to maximize accumulated reward, typically by introducing a sampling cost. Therefore, by fixating largely on the sequential nature of the tasks, these studies focus only on a particular efficiency–performance tradeoff known as the speed–accuracy tradeoff (Del Giudice and Crespi, 2018) In many complex decisions, however, there are several other functional tradeoffs that involve other properties of the agent–environment loop, such as limited sampling resources, limited interactions with the environment and delayed feedback (Moreno-Bote et al., 2020). The effect of these resource limitations on decision making might not be important when there are only two or three available options, but it might be critical when going beyond those low numbers. In that case, the allocation of resources might be governed by two-stage processes (Hauser and Wernerfelt, 1990; Mehta et al., 2003; Shocker et al., 1991; Roberts and Lattin, 1991), instead of purely sequential processes, where the first decision is about the subset of options that will be considered for further processing.

Here we study whether narrowing attention to a few options results from optimally allocating finite resources in multi-alternative choices. To this end, we consider an infinitely divisible sampling resource (e.g. time or precision), such that there are no bounds in the number of alternatives that can be considered. In our model, an agent can first allocate finite sampling time over an arbitrarily large number of options, modelled as accumulators of noisy evidence, with the only restriction that the total sampling time is fixed. This is in stark contrast with previous work on the breadth–depth tradeoff, where the sampling process was simplified, and where the sampling outcomes and resources were discrete, thus obtaining qualitatively different predictions (Moreno-Bote et al., 2020). This accumulation of evidence runs in parallel and independently for each accumulator, and only their final states are observed. Based on the observations, the agent picks up the one with the highest expected rate of evidence accumulation, which defines the utility of the choice. The goal of the agent is to optimize the allocation of sampling time such that expected utility is maximized. We identify a critical variable in the problem, that we simply call *capacity*, that increases with the actual size of the resources of the agent as well as with the discriminability between options, and we find that this capacity separates two distinct regimes of optimal allocation. When sampling capacity is small, the optimal policy is to sample exactly five options, regardless of the prior. In contrast, when capacity is large, the number of options to sample grows with capacity in a sub-linear fashion that depends on the prior. We find a duality between allocated time and allocated precision to the options, such that all our results generalize to allocating precision while keeping fixed sampling time. Finally, we show that even allocations are optimal, and thus better than more complex asymmetric time allocations over the considered options. Overall, our results suggest that decisional bottlenecks can be a byproduct of optimal policies in the face of uncertainty.

3.2 Multi-accumulator model

We consider an environment that generates many options ($N \gg 1$) from which to choose (Fig. 1, top), each one characterized by a 'drift' parameter μ_i ($i = 1, \ldots, N$), unknown to the agent. All drifts μ_i are drawn identically and independently from a prior probability distribution $p_\theta(\mu)$, known to the agent and assumed to have finite mean and variance. In order to choose between the options, the agent gathers information by sampling them. The critical aspect of our model is that sampling times $t_i \geq 0$ (Fig. 1, bottom) need to be allocated before the actual sampling occurs, and with the constraint that the total sampling time T is limited,

$$\sum_{i=1}^{N} t_i = T. \tag{3.1}$$

In practice, the agent needs to decide on the number of options $M \leq N$ to be sampled and their corresponding sampling times $t_i > 0$ for $i \leq M$, while the remaining options $i > M$ are ignored by giving them no sampling time, $t_i = 0$ (Figure 1, bottom). The ordering of the options is irrelevant, as they are initially indistinguishable, and thus we take the first M as those that are sampled. We assume that non-sampled options cannot be chosen, although a 'default' option can be added to our framework with no change of our main results.

Once total sampling time is allocated, noisy evidence about the drift μ_i of each of the sampled options $i \leq M$ is integrated by independent accumulators (Fig. 1, middle) according to the drift-diffusion process

$$\frac{dx_i(t)}{dt} = \mu_i + \eta_i(t), \tag{3.2}$$

where $x_i(t)$ is the accumulated evidence up to time t with initial condition $x_i(0) = 0$, and $\eta_i(t)$ is a Gaussian white noise with zero mean and fixed variance σ^2, independent and identical for all the accumulators.

The result of the accumulation is the total evidence x_i at time t_i, both of which are observed by the agent and constitute the sufficient statistics for the unknown drift μ_i (Moreno-Bote, 2010). With these observations, the agent builds the posterior distribution of the drifts by using Bayes rule as

$$p(\mu_i|x_i, t_i, \sigma, \theta) = \frac{\mathcal{L}(\mu_i|x_i, t_i, \sigma)p_\theta(\mu)}{p(x_i|t_i, \sigma, \theta)}, \tag{3.3}$$

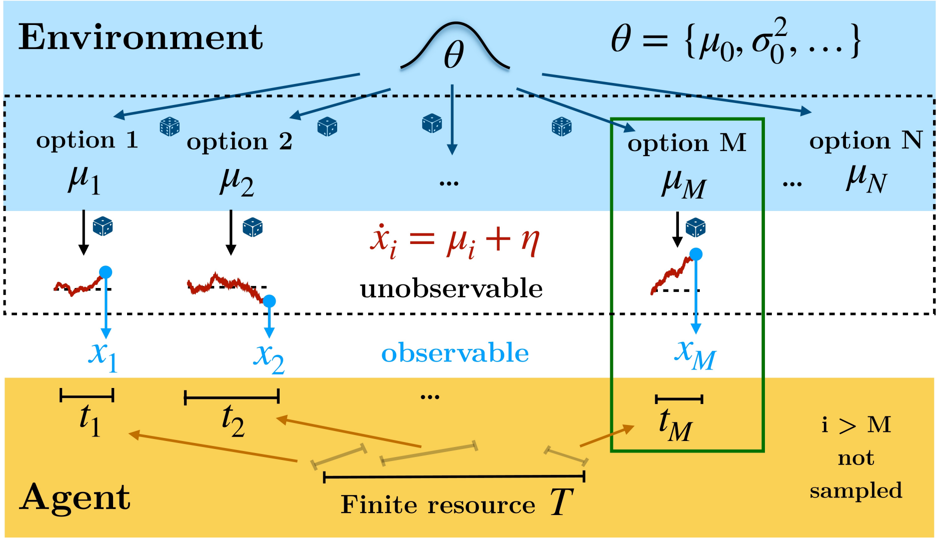


Figure 3.1: A multi-accumulator model with finite sampling resources. The environment produces a large number of options, each characterized by a drift μ_i, unknown to the agent and drawn from a prior distribution characterized by hyperparameters θ, which is known to the agent. The agent has a finite resource T, that they divide and allocate across options, $\sum_i t_i = T$, in order to sample them. In practice, the agent allocates finite sampling time to a finite number M of accumulators to infer their unknown drifts. After allocation, evidence (red lines) is optimally integrated by the accumulators. The agent observes the integrated evidence x_i at the end of the accumulation, after time t_i, infers the drifts for each of the accumulators and chooses the one that is deemed to have the highest drift (in this case, μ_M; green box).

where $\mathcal{L}(\mu_i|x_i, t_i, \sigma) = \mathcal{N}(x_i|\mu_i t_i, \sigma^2 t_i)$ is the likelihood function for the drift, $p_\theta(\mu)$ is the prior distribution and $p(x_i|t_i, \sigma, \theta) = \int \mathrm{d}\mu \, \mathcal{N}(x_i|\mu t_i, \sigma^2 t_i) p_\theta(\mu)$ is the marginal distribution of the evidence, which serves as a normalization constant.

After building these posterior distributions, the agent simply chooses the option with the highest expected drift, (Fig. 1, middle, green box), which defines utility, $U(M, \mathbf{x}, \mathbf{t}, \sigma, \theta) = \max_{i \le M} \hat{\mu}_i (x_i, t_i, \sigma, \theta)$, where $\mathbf{x} = (x_1, ..., x_M)$ is the vector of observations for the M accumulators with allocated times $\mathbf{t} = (t_1, ..., t_M)$. To avoid notation clutter, from now on we will stop writing the dependence on σ and θ of the various functions and leave it implicit.

The previous expression is the utility of the choice of accumulator, which depends on the observations and allocation times. However, before time is allocated, the observations $\mathbf{x}$ themselves will be unknown to the agent. Therefore, the expected utility of a given allocation $\mathbf{t}$ is given by taking the expectation of the above utility over all possible observations as

$$\hat{U}(M,\mathbf{t}) \equiv \mathbb{E}\left[\max_{i\leq M}\hat{\mu}_i|\mathbf{t}\right] = \int \mathrm{d}x_1 \ldots \mathrm{d}x_M\, p(x_1,\ldots,x_M|\mathbf{t}) \max_i \hat{\mu}_i\left(x_i,t_i\right), \quad (3.4)$$

where, using the independence of the accumulators, $p(x_1,...,x_M|\mathbf{t}) = \prod_{k\leq M} p(x_k|t_k)$ is the product of the marginal distribution of the evidences.

Optimally inferring the drifts from observations is readily accessible through Bayesian inference as shown above. Thus, the main, and harder, objective of the agent is to optimize the allocation policy, i.e. to select both the number of sampled accumulators $M \leq N$ and the time t_i allocated to each, in order to maximize expected reward, while satisfying the total sampling time constraint in Eq. (3.1). This is accomplished by optimizing the utility with respect to M and $\mathbf{t} = (t_1,...,t_M)$ as

$$(M^*,\mathbf{t}^*) = \arg\max_{M,\,\mathbf{t}} \hat{U}(M,\mathbf{t}). \quad (3.5)$$

3.3 Capacity and time–precision duality

While time can be understood as the resource that the agent allocates, we found a dimensionless scale that expresses their actual sampling capacity, i.e. their ability to sample and differentiate between drifts, which we call capacity C (Fig. 3.2). As the agent integrates noisy evidence through Eq. (3.2), the likelihood of the drift μ_i for accumulator i is proportional to a Gaussian (Fig. 3.2**a**, orange curve) with mean x_i/t_i and variance σ^2/t_i, $\mathcal{L}(\mu_i|x_i,t_i,\sigma) \propto \mathcal{N}(\mu_i|\frac{x_i}{t_i}\frac{\sigma^2}{t_i})$ (Moreno-Bote, 2010). Its variance σ^2/t_i shows how the sampling time and the variance of the sampling noise are related when inferring the drift μ_i: the likelihood gets broader with increasing σ or decreasing time t_i, reflecting that the precision of the observations is decreased by having more noise or less time, respectively. In fact, the sampling capacity of the agent should capture this duality. Thus, having a fixed capacity could be interpreted as having a fixed noise variance σ^2 for all accumulators and allocating time T between them (Fig. 3.2**b**, left) or as having a fixed sampling time T for each of the accumulators and allocating precision $1/\sigma^2$ between them (Fig. 3.2**b**, right).

Moreover, the posterior in Eq. (3.3) depends on the prior as well (Fig. 3.2**a**, cyan curve). For fixed evidence, the broader the prior is, the easier it is to differentiate between sampled drifts, since the expected squared distance between two drifts drawn from the same distribution is twice its variance $\mathrm{Var}[p_\theta(\mu)]$ (see Appendix 3.6.1). Therefore, we define the capacity allocated to option i as the ratio between the precision of the observation and the precision of the prior,

$$c_i = \frac{\mathrm{Var}[p_\theta(\mu)]}{\mathrm{Var}[\mathcal{N}(\mu_i|\frac{x_i}{t_i},\frac{\sigma^2}{t_i})]} = \frac{\sigma_0^2}{\sigma^2}t_i. \quad (3.6)$$

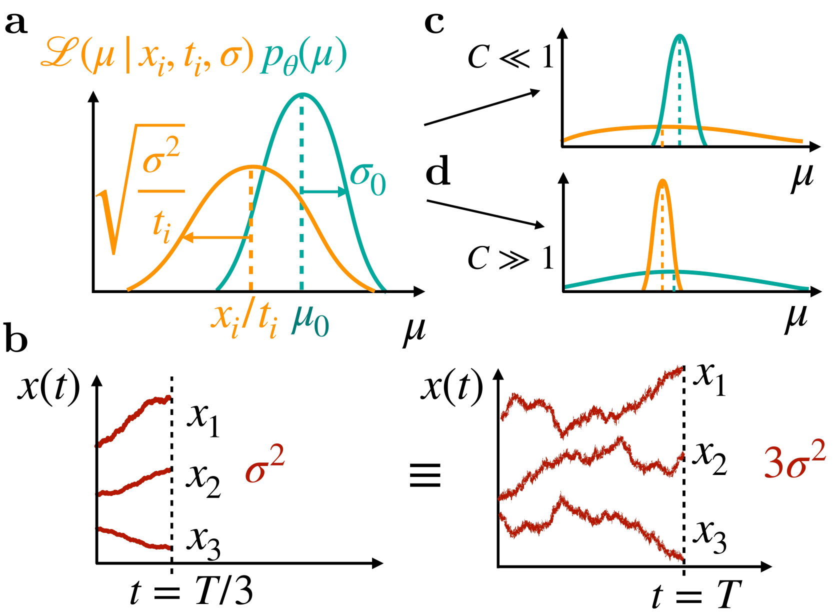


Figure 3.2: Time/precision duality and the notion of capacity. (**a**) The likelihood of the drift μ (in orange) given the evidence has variance σ^2/t_i and the prior distribution of the drifts (in cyan) has variance σ_0^2. These quantities determine capacity as in Eq. (3.6). (**b**) Time and sampling noise are intricately related (see text). In this example, allocating time $T/3$ to each accumulator under fixed precision $1/\sigma^2$ (left) is equivalent to allocating precision $1/3\sigma^2$ to each accumulator under fixed sampling time T (right). (**c**) Small capacity means that the variance of the observation is much larger than the variance of the prior, indicating that it is difficult to confidently identify the best drift from the observations. (**d**) In the large capacity limit, it is easier to differentiate the good drifts from the poor ones.

Adding the individual capacities results in the total sampling capacity of the agent,

$$C = \sum_i c_i = \frac{\sigma_0^2}{\sigma^2} T. \tag{3.7}$$

For the rest of this article, we stick to the interpretation of allocating capacity as dividing the total time T while fixing the accumulation noise σ, such that the variable we can control is the sampling time allocated to each option, keeping in mind that all the results presented below can be readily reinterpreted as dividing precision while giving to all options the same sampling time.

3.4 Results

Optimally dividing sampling capacity C into options is an *a priori* hard problem due to its high dimensionality. However, we show in subsection 3.4.2 that the optimal allocation lies within the family of even allocations, where M options receive equal sampling time $t_i = t \equiv T/M$, while the remaining others are given no time. Thus, finding the optimal policy reduces to finding the optimal number M of accumulators to sample.

3.4.1 Even sampling

In this subsection, we exploit the structure of even sampling. First, the posterior mean of the drift $\hat{\mu}_i(x_i, t)$, computed from Eq. (3.3), is a monotonously increasing function of the evidence x_i for any prior (see proof in Sec. 3.6.2 in Appendix). Therefore, the option that maximizes the posterior mean $\hat{\mu}_i$ is the one that has the highest evidence $x_i(t)$, as all M sampled options are given the same sampling time t. This allows us to work by maximizing evidence instead of maximizing the posterior means of the drifts in Eq. (3.4). Secondly, by changing variables $y \equiv \max_i x_i$, and using the probability distribution of the maximum y, denoted $p_{\max}(y|t, \sigma, \theta)$, the expected utility in Eq. (3.4) can be recast in the one-dimensional integral

$$\hat{U}(M, t) = \int \mathrm{d}y \; p_{\max}(y|t) \hat{\mu}(y, t).$$

Finally, given that the M options are sampled evenly, the probability distribution of the maximum can be simplified by using the cumulative distribution of the evidence x for an arbitrary accumulator, $F_x(y|t) = \int_{-\infty}^{y} \mathrm{d}x' \, p(x'|t)$, where $p(x|t)$ is the marginal of the evidence x of the accumulator, as

$$p_{\max}(y|t) = \frac{\mathrm{d}}{\mathrm{d}y} \left[F_x(y|t)\right]^M . \tag{3.8}$$

With all the above, the expected utility in Eq. (3.4) can thus be written as

$$\hat{U}(M, t) = M \int \mathrm{d}y \; \left[F_x(y|t)\right]^{M-1} p(y|t) \, \hat{\mu}(y, t). \tag{3.9}$$

When the prior distribution is a Gaussian with mean μ_0 and variance σ_0^2, it is possible to identify the total capacity $C = \frac{\sigma_0^2}{\sigma^2} T$ explicitly and Eq. (3.9) simplifies to

$$\hat{U}(M, C) = \mu_0 + \frac{M\sigma_0}{\sqrt{1 + \frac{M}{C}}} \int_{-\infty}^{\infty} \mathrm{d}y \, \left[\Phi(y)\right]^{M-1} \mathcal{N}(y|0, 1) \, y, \tag{3.10}$$

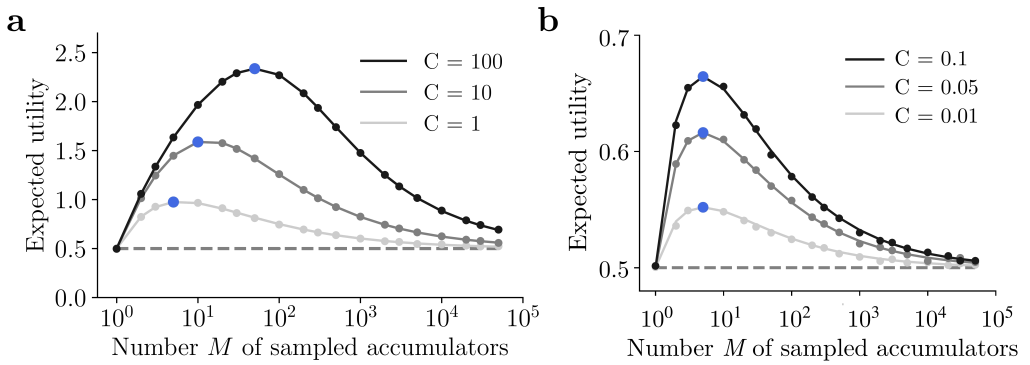


Figure 3.3: Expected utility as a function of sampled accumulators exhibits the breadth-depth tradeoff. Results for the Gaussian prior case ($\mu_0 = 0.5, \sigma_0^2 = 1$), for various different capacities. Blue points denote the maxima. Note log horizontal scale (points, Monte Carlo simulations; lines, theoretical predictions, Eq. 3.10). (**a**) For large capacities, the optimal number of sampled accumulators changes with the capacity. (**b**) For small capacities, the optimum is independent of capacity and equal to five.

where $\Phi(y) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y}{\sqrt{2}}\right)\right]$ is the cumulative distribution function of a normal distribution.

Plotting the utility in Eq. (3.10) as a function of the number of sampled accumulators M reveals a clear breadth-depth tradeoff (Fig. 3.3a). At the depth limit, $M = 1$, only one accumulator is sampled and it is given all sampling time T. In this case, the expected utility will simply be the expected value of the prior, $\mu_0 = 0.5$ (Fig. 3.3a, left point), since there is no choice to be made between accumulators. At the breadth extreme, $M/C \to \infty$, the evidence gathered for each accumulator is very noisy because each has been allocated a very short sampling time, and thus choosing any will amount to an expected utility again equal to the prior mean (rightmost points). Therefore, for all capacities, there is an intermediate optimal value for the number of accumulators to sample, M^*.

Sharp transition between the small and large capacity regimes

Our main result is that the optimal allocation policies are qualitatively different at small and large capacity, and that there is a abrupt transition between the two regimes. We provide useful asymptotic analytical expressions for the utility in Eq. (3.9) and the optimal M^* in both limits and describe their characteristic features.

The limit $C \ll 1$ corresponds to the case where the uncertainty in the observation σ^2/T is much larger than the variance of the prior σ_0^2, i.e. the Gaussian likelihood is much wider than the prior (Fig. 3.2**c**). In this limit, we find that the utility in Eq. (3.9) can be expanded a series in powers of $\sqrt{C}$, which at first order is given by (see Appendix, Sec. 3.6.3)

$$\hat{U}(M,C) \approx \mu_0 + \sigma_0 \sqrt{\frac{C}{2\pi}} \left[\sqrt{M} \int_{-\infty}^{\infty} \mathrm{d}z \, z \exp\left(-\frac{z^2}{2}\right) \left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right)^{M-1} \right] \tag{3.11}$$

Remarkably, this expression holds for any prior distribution as long as capacity is small enough. Let us now note that the only dependence on M appears in the quantity in square brackets, so we isolate it to look for the optimal M. Since capacity does not play a role here, we see that M^* will be constant as a function of small capacity. Furthermore, using Extreme Value Theory (see Sec. 3.6.5 in Appendix), we find that this quantity decreases with M for large M. This means that sampling many accumulators will not be optimal, following the intuition that there is no point in sampling many options when having scarce resources. On the other hand, and as noted before, it is easy to see that expected utility attains its lowest value when $M = 1$, since in this case there is no choice to be made. Thus, the optimal M is attained at some intermediate value. Given these observations, the optimum can be thus found numerically by varying M, and it happens when

$$M^*(C \ll 1) = 5,$$

which can be checked visually for various values of small capacity in Fig. 3.3**b**, which also validates the approximation in Eq. (3.11). It is important to highlight that when capacity is strictly zero, expected utility does not depend on M, and is equal to the prior mean, since choosing options to sample has no effect when there is no time to be allocated. However, as long as the small capacity is finite, an optimal number of options to sample equal to five emerges, regardless of the prior and the value of capacity. We have confirmed this strong prediction by direct numerical integration of Eq. (3.9) using different prior distributions, including Gaussian, uniform and bimodal (Fig. 3.4), which also holds even when a non-sampled, default, option can be chosen (diamond markers).

The opposite limit $C \gg 1$ corresponds to the case where the precision of the observation is much greater than the one of the prior (Fig. 3.2**d**). Intuitively, this means that the quality of the observations is good enough to likely differentiate the drifts between two randomly chosen accumulators, and thus we expect the optimal number of accumulators to increase with increasing capacity, giving a qualitatively different behaviour than at small capacity. Hence, we make this assumption to inspect the optimality of Eq. (3.9) for this large capacity limit, which we find to be consistent with the numerical results shown below. In particular, when the prior distribution is Gaussian, the expected utility in Eq. (3.10) has the following asymptotic behavior,

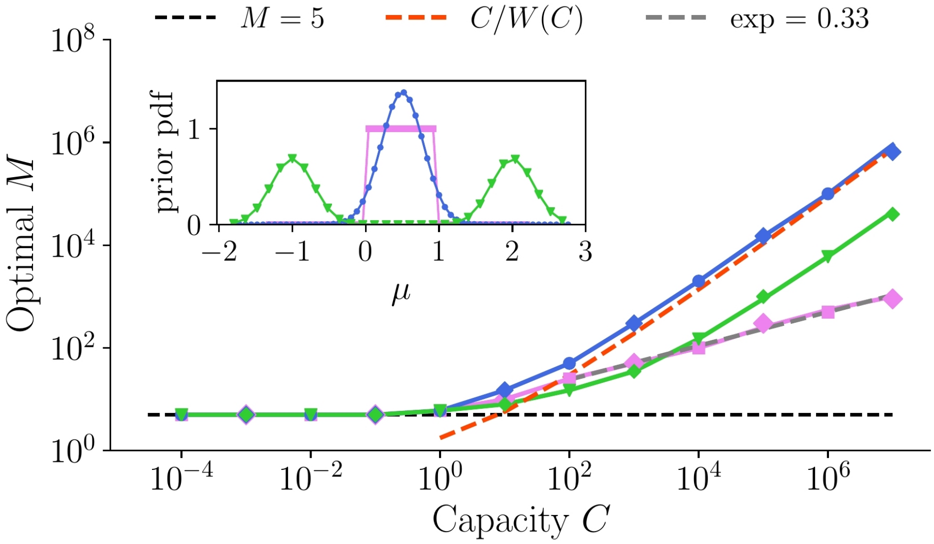


Figure 3.4: The optimal number of sampled accumulators undergoes qualitatively different behaviors at small and large capacity values. Results come from searching the maximum expected utility via Monte Carlo simulations (points) and numerical integration (lines) for a Gaussian (blue line; Eq. 3.10), uniform (pink; Eq. 3.20) and a bimodal (green; Eq. 3.21) priors (illustrated in inset). We used $\mu_0 = 0.5$ for all priors and $\sigma_0^2 = 1/12$ for the Gaussian prior to match the uniform distribution. For the bimodal prior, the variance of each mode equals σ_0^2. Dashed red line corresponds to the asymptotic limit in the Gaussian prior case, Eq. (3.13). Dashed gray line (almost overlaid by the pink line) is the best power law fit for the uniform prior case ($M^* \propto C^a, a = 0.33$). Diamonds (overlaying most of points) indicate simulations with a 'default' option.

$$\hat{U}(M \gg 1, C) \to \mu_0 + \sigma_0 \frac{b_M}{\sqrt{1 + \frac{M}{C}}}, \tag{3.12}$$

where $b_M = (2\log(M) - \log(\log(M)) - \log(4\pi))^{1/2}$ (see Sec. 3.6.5 in Appendix). By relaxing M to be continuous, we can maximize expected utility, and we find that the optimal number of sampled options for large capacity satisfies, up to leading order, the implicit equation $M^* \log(M^*) = C$. After inverting it, the optimal number of sampled options is

$$M^*(C \gg 1) = \frac{C}{W(C)}, \tag{3.13}$$

where $W(C)$ is the Lambert function. This asymptotic limit provides a very good approximation to the optimal M^* at large C obtained from direct numerical integration of Eq. (3.10) (Fig. 3.4; red dashed line, theory; blue points, simulations). For prior distributions other than the Gaussian, we rely on numerical integration of Eq. (3.9) (see Secs. 3.6.6 and 3.6.7 in Appendix for analytical expressions). For a uniform prior, the optimal number of sampled options increases as a power law with an exponent close to $1/3$ (Fig. 3.4, pink), while for a bimodal prior the optimal number increases in a similar fashion to the Gaussian prior case (green). While differences of asymptotic limits are due to the presence of bounded or unbounded drifts in the priors, in all cases the increase is sub-linear, indicating that increasingly longer times are allocated to each of the sampled accumulators as capacity increases.

The above results show that there are two distinct regimes, one at small and another at large capacities, characterized by qualitatively different optimal allocations: while at small capacity the optimal number of sampled accumulators should be five regardless of the prior, at large capacity the optimal number of sampled accumulators grows sublinearly regardless of the tested prior. Further, we observe that there is an abrupt transition between the two regimes as capacity grows, with a bump being observed at intermediate capacity values.

3.4.2 Even allocation is optimal

Above we have assumed that we could find the optimal time allocation within the subset of even allocations, such that, given finite total time T, an agent just needs to determine how many options will be sampled and split equal time to all of them. Conveniently, this set is discrete and thus amenable to effective search of the optimum. However, in general, the set of allocation policies is the infinite-dimensional simplex $\sum_i t_i = T, t_i \geq 0$ for all i, as a priori the agent could unevenly split time to options in any arbitrary way. Despite its infinite-dimensionality, we have seen in the case of even sampling that it is optimal to ignore (infinitely) many options, such that $t_i > 0$ only for $i \in \{1, ..., M\}$, with finite M, to which we will refer as having M active dimensions.

To address the most general case, using the above intuitions we first generalize the expected utility, Eq. (3.9), to the case when allocated time is unevenly distributed among M accumulators, as

$$\hat{U}(M, \mathbf{t}) = \int_{-\infty}^{\infty} \mathrm{d}y \, \frac{\mathrm{d}}{\mathrm{d}y} \left[\prod_{i=1}^{M} F_x(y|t_i) \right] \hat{\mu}(y, t_i), \tag{3.14}$$

where $F_x(y|t_i)$ is the cumulative distribution function of the posterior when using t_i sampling time. Our goal is then, for every M, to find the allocation $\mathbf{t}$ that maximizes Eq. (3.14) under the capacity equality constraint and the inequalities

$t_i \geq 0$ for all i, and then select the optimal M, the one that achieves the highest utility.

In this more general setup, an even allocation corresponds to the symmetrical point in M active dimensions given by $\mathbf{t}^e_M$, where $t^e_{M,i} = T/M$ for $i = 1, ..., M$ (superscript reflects 'even' allocation). As the expected utility in Eq. (3.14) is symmetric under any permutation $t_j \leftrightarrow t_k$ for any j and k, all its partial derivatives have to be equal at $\mathbf{t}^e_M$. Therefore, every even allocation for each M corresponds to a critical point of the constrained optimization problem (see Appendix, Sec. 3.6.8).

We still need to characterize these critical points in order to show that the global maximum is indeed an even allocation. We first remember that the optimal number of active dimensions M needs to be found, and thus it is useful to see how expected utility varies as a function of M. To do this, we note that any M-dimensional simplex is in fact the border of an $(M + 1)$-dimensional simplex. For example, for $M = 2$, the constraints describe a line segment, or 1-simplex, where we have the symmetric critical point $\mathbf{t}^e_2 = (T/2, T/2)$ (Fig. 3.5**b**, black circle). We then notice that the line $t_1 + t_2 = T$ is one of the 3 edges of the triangle, or 2-simplex (Fig. 3.5**c**: pink lines are the edges of triangle), where in fact we have another symmetric critical point in its interior (black triangle). With this, we can 'visualize' the infinite-dimensional nature of this problem, since all critical points of the utility lie at the edges of a higher dimensional simplex.

To asses the landscape of expected utility in high-dimensional simplices, we can evaluate it at all symmetric critical points $\mathbf{t}^e_M$ and along directions that go orthogonally between them (Fig. 3.5**c**, orange arrows). Thus, we devised a one-dimensional path that allows to continuously connect all symmetrical critical points, and applied it to the small capacity limit $C \ll 1$. As we move from the 1-simplex to higher dimensional simplices (as in Fig. 3.5**c**), we find that first utility increases, reaching a maximum at the even allocation in $M = 5$ dimensions, and then decreases (Fig. 3.5**d**). Therefore, critical points $\mathbf{t}^e_2$, $\mathbf{t}^e_3$ and $\mathbf{t}^e_4$ are 'saddle'-like points, as they are maxima in the interior of their corresponding simplex, and minima as one moves to the interior of the higher dimensional simplex.

Although the above analysis suggests that the optimum lies at an even allocation point, it is still unclear whether there are other critical points that are asymmetrical and have a larger utility. To argue that the presence of non-symmetrical local optima is unlikely, we used a stochastic gradient projection method (Fletcher, 2013) that maximizes expected utility subject to the constraints, and applied it to the Gaussian prior case (see Sec. 3.6.8 in Appendix for details). Indeed, we find for various capacities that, regardless of the initial condition, i.e. random initial allocations, a maximum utility is attained when time is evenly divided (Fig. 3.5**e**), and the global maxima coincide with the ones found in the previous sections.

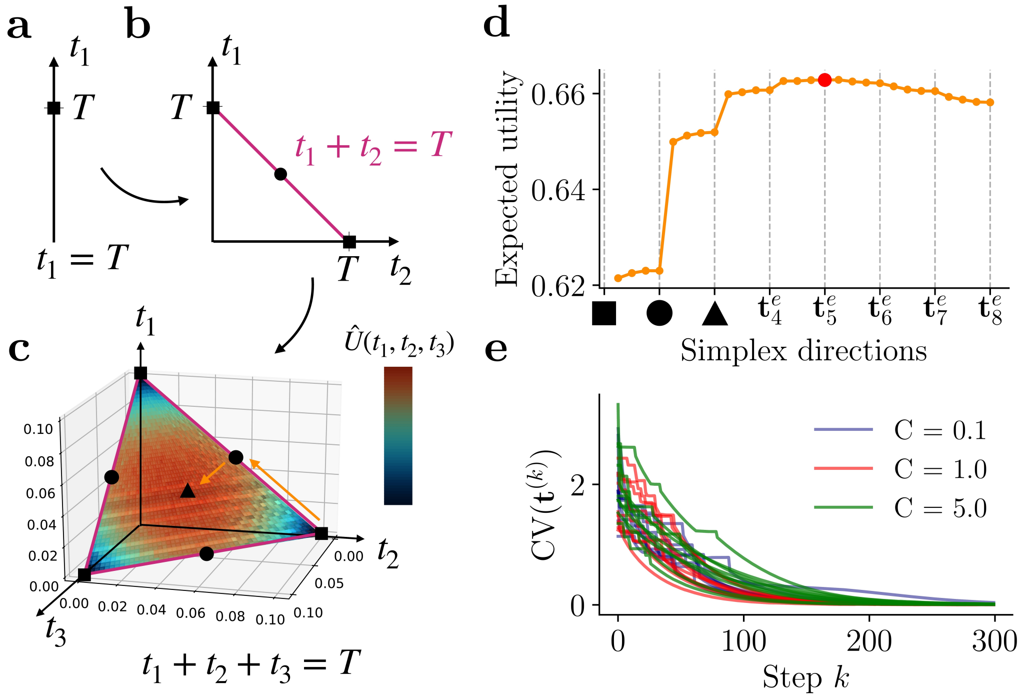


Figure 3.5: Even allocations correspond to critical points of utility lying at the center of M-simplices. **(a)** In one dimension, there is only one point that complies with the constraint. **(b)** For $M = 2$ dimensions, constraints define a line segment or 1-simplex. The circle depicts the symmetric critical point $\mathbf{t}_2^e$. **(c)** For $M = 3$, constraints form a triangle or 2-simplex. The black triangle is the symmetric critical point $\mathbf{t}_3^e$. The colors at the extremes reflect the minimum and maximum utility reached in this simplex, which was computed with Monte Carlo simulations of Eq. (3.4) for the Gaussian prior with $T = 0.1$, $\sigma = 1$, $\sigma_0 = 1$, $\mu_0 = 0.5$. **(d)** Expected utility computed along directions that go orthogonally from $\mathbf{t}_M^e$ to $\mathbf{t}_{M+1}^e$ (as illustrated with orange arrows in panel **c**, same parameters). The red dot shows the maximum occurring at $\mathbf{t}_5^e$. **(e)** Using the stochastic projected gradient ascent detailed in Sec. 3.6.8, we initialized the algorithm at random points (ten shown here) in a high-dimensional simplex and measured the coefficient of variation (CV) of the allocation vector at every step of the algorithm until convergence, for various values of capacity. Zero CV implies even allocation.

3.5 Discussion

We have studied a model of multi-alternative decision making where an agent can allocate finite sampling resources to options and choose the best one amongst them. We found that the capacity of the agent depends on both the amount of sampling resources, i.e. time or precision, as well as on the discriminability of the options in the environment. As a function of capacity, optimal policies undergo an abrupt transition: at small capacity, allocating time to a handful of options is optimal; at large capacity, the number of options grows sub-linearly, well below the actual sampling capacity of the agent. Our results show that decision bottlenecks, such as option-narrowing, can arise from optimal policies in the face of uncertainty, and provide so far untested predictions on choice behaviors in multi-alternative decision making as a function of capacity.

Seemingly strict limits pervade cognition, from the so-called attentional bottleneck (Deutsch and Deutsch, 1963; Treisman, 1969; Yantis and Johnston, 1990), over working memory (Miller, 1956; Cowan et al., 2005; Luck and Vogel, 2013; Ma et al., 2014; Brady et al., 2011), to executive control (Shenhav et al., 2017; Norman and Shallice, 1986; Sleezer et al., 2016). These limits might result from using scarce neuronal resources or from using them inefficiently. However, a likely alternative is that bottlenecks reflect strategies that make optimal use of limited but large resources. Indeed, past work has recognized that some apparent limits, most notably dual tasking bottlenecks (Fischer and Plessow, 2015; Meyer and Kieras, 1997), could be the result of optimal allocation of finite resources to avoid overlap and interference between the different representations needed to solve the two tasks (Meyer and Kieras, 1997; Feng et al., 2014; Zylberberg et al., 2011). Further, it has been recognized that the narrow focus of attention could be at the heart of solution to the the binding problem by integrating separate features into a coherent object (Treisman, 1998), and thus its narrowness might reflect a function more than a limitation. Our work follows this line of argument and provides for the first time a quantitative account for why it is optimal for an agent to consider a handful of options in the face of uncertainty, well above two but well below 10. In addition, our results shed light on why people might ignore hundreds of accessible options and focus resources to a very small number of options (Hauser and Wernerfelt, 1990; Iyengar and Lepper, 2000; Scheibehenne et al., 2010). Thus, some of the seemingly strict limits in decision making can be the result of optimal policies that favour depth versus breadth processing of the options.

It has been long recognized that people often consider a small set of options while ignoring many others (Stigler, 1961; Hauser and Wernerfelt, 1990; Mehta et al., 2003; Payne, 1976). In the 'consumer' literature this is explained by arguing that small consideration sets are favored because they optimally balance the probability of finding a good option in the set with the search and mental costs incurred in adding new options to that set. These models thus assume that resources are not limited, but are costly. In contrast, the assumptions in our work do not explicitly tune the cost of sampling, but rather an implicit cost arises naturally from the strict capacity constraint, which depends intrinsically on the agent as well as extrinsically on the environment. A more fundamental distinction is that previous work did not focus on allocating resources intensively into the options, such that the only decision was whether to include an option into the set or not, without considering the amount of resources allocated to it. This distinction makes that problem drastically different than the tradeoffs of the breadth-depth dilemma considered here. This can explain why transitions of optimal policies as a function of agent's parameters have not been reported before.

Most current theories of perceptual and value-based decision making are based on accumulation of evidence that favors certain hypotheses over others (Gold and Shadlen, 2007; Drugowitsch et al., 2012; Moreno-Bote, 2010; Ratcliff and Smith, 2004). When combined with sampling costs and rewards, a normative theory emerges where it is optimal to only accumulate evidence up to a bound (Gold and Shadlen, 2007; Drugowitsch et al., 2012). By letting the information between the competing options be processed parallelly, a tradeoff between speed and accuracy of the choice emerges, and it is then possible to derive optimal policies under various further assumptions (Vul et al., 2014; Tajima et al., 2019; Callaway et al., 2021; Jang et al., 2021). For instance, the work of Vul and colleagues assumes that only one option is correct, that is, that there are only two types of reward. Combined with an opportunity cost, taking zero to one sample in this setting becomes optimal under large parameters regions of the cost of time (Vul et al., 2014). However, multi-alternative decision making requires estimating the subjective value of offers, and none of them is correct in any absolute sense, which can favor not single-sample, deep strategies (Moreno-Bote et al., 2020). The work of Jang and colleagues aims to optimally allocate attentional resources by solving numerically the Bellman equation in binary choice (Jang et al., 2021), but extending this framework to (many) multi-alternative decisions is intractable. While the speed-accuracy tradeoff is a ubiquitous phenomenon in sequential decisions, there are certainly other features of decision making that need to be studied in isolation to advance our understanding of the various challenges that arise in multi-alternative decision making.

In our work, we have implemented the prevalent feature of evidence accumulation, but we have highlighted other characteristics of decision making, such as finite sampling resources, delayed feedback and limited interaction with the environment, that are critical in many real-world examples (see Introduction). These assumptions differentiate our results with those of previous work. Most importantly, we have not considered a sequential process where the evidence gathered during the accumulation is observable and thus it cannot be used to stop the accumulation process.

This is important case when accumulation of evidence happens in a decentralized manner by, e.g., different groups of neurons, and only the final result of the accumulation is conveyed to another set of neurons where the comparison and choice takes place. Indeed, parallel sampling of information is a ubiquitous ingredient in theories of decision making (Busemeyer and Townsend, 1993; Glöckner and Betsch, 2008). In this work, limited capacity in parallel sampling is understood as a limitation on available sampling precision, thus allowing for the allocation of attentional resources under parallel evidence accumulation with time pressure (see Figure 2). These ingredients force the decision maker, in a deliberate or automatic way, to allocate resources into the options in advance in a strategic fashion, prompting the need to trade sampling breadth over depth.

Although previous work has characterized optimal breadth-depth tradeoffs in multi-alternative choices like the ones studied here, it has been assumed that agents have a finite 'discrete' capacity (Moreno-Bote et al., 2020). Our assumption of a continuous resource (e.g. time or precision) that can be infinitely divided has allowed us to uncover qualitatively novel optimal policies at small capacity. We have therefore been able to derive optimal policies that trade off breadth with depth search in (many) multi-alternative settings where using traditional sequential decision frameworks would be intractable. Integrating resource allocation with sequential decision making into a single theory of dynamic allocation will be most relevant to understand human decision making, but its study is deferred to the near future. In any event, any agent with finite capacity cannot avoid the problem of first deciding how many options to allocate capacity to, and thus breadth-depth tradeoffs as described above will be generally at play.

Bounded rationality accounts (Simon, 1972; Russell and Wefald, 1991; Gershman et al., 2015; Griffiths et al., 2015) propose that cognition results from the finite limits of the nervous system from where it emerges. Our work follows this line of research in two ways. First, we propose that agents indeed have a finite sampling capacity that can be arbitrarily allocated to the available options. However, an important assumption in our work is that while the intrinsic resources of an agent might seem large, the interaction of the agent with the environment might render their effective decision-making capacity small. Therefore, capacity is not an absolute quantity that describes an agent, but a relative quantity that contextualizes the agents and characterizes how well they are suited to solve a given task in the world. An important contribution of our work is to show that optimal policies depend on effective capacity in a highly non-linear way, such that small-capacity agents would behave qualitatively different than large-capacity agents (or even the behavior of the same agent operating in different environments could be qualitatively different). This is clearly a prediction that can be tested with humans where time or other resources are constrained and varied on a trial by trial basis. Secondly, agents perform the allocation before feedback is received, which relates to a bounded-optimal agent that is optimized at 'design'-time, which eliminates the paradox of perfect rationality by not letting the agent optimize their decisions at run-time (Russell and Subramanian, 1994), an argument that further supports the validity and relevance of two-stage decisions.

Another important result of our work is that evenly dividing time to a small set of options is optimal when they are initially indistinguishable. This optimal division of resources coincides with the $1/N$ heuristic rule (Gigerenzer and Gaissmaier, 2011) or equality heuristic (Messick, 1993), which has proven to be implemented in human decision making and highly efficient as a portfolio strategy (DeMiguel et al., 2009). In our case, the fact that options are drawn from the same prior (known to the agent) contributes to the optimality of the even allocation. Although the optimal allocation of non-identically distributed options is not addressed here, this heuristic can be efficient in such situations (Thorngate, 1980). It is important to realize that the optimal low numbers of considered options have been found in the case where their values are not known in advance and come from the same distribution. If agents have strong preferences or have additional information about the expected values of the options (e.g. by sampling them sequentially), then the number of considered alternatives will be further reduced. Of course, if the agents are allowed to sequentially sample options with which they are familiar, a non-even allocation might emerge to be optimal (Callaway et al., 2021; Tajima et al., 2019; Sepulveda et al., 2020). Nonetheless, for binary choice, reward is still maximized at even allocations in sequential sampling when options have not been unevenly sampled in the past (Fudenberg et al., 2018; Jang et al., 2021). On the other hand, when the number of alternatives is much higher than two, people choose to ignore many of the available options (Thomas et al., 2021), consistent with our findings. Moreover, for fixed-duration tasks there is evidence that humans have a choice set of around five in sequential decisions (Reutskaja et al., 2011), even if their final allocation might be uneven. This shows once again that a low number of considered options can hardly be taken as evidence of a decisional bottleneck and is more in line with an optimal tradeoff between breadth and depth.

Finally, our results can have important implications for the optimal wiring of neural networks in the brain (Rushworth et al., 2011; Siegel et al., 2015; Vickery et al., 2011; Yoo and Hayden, 2018). First, as just few options should be considered at the same time, it is expected that only those would be encoded in different, albeit possibly overlapping, pools of neurons. Thus, although models consisting of two or three pools that compete for dominance through mutual inhibition can be a sensible idea for binary and ternary decision making (Gold and Shadlen, 2007; Cisek and Kalaska, 2010; Roe et al., 2001; Usher and McClelland, 2001; Moreno-Bote et al., 2007; Churchland et al., 2008; Wang, 2008), extrapolating this to many more options (e.g., larger than 10) by splitting neurons into corresponding pools of neurons would be hardly optimal. Our results are, in contrast, consistent with the opposite view that posits that a single pool of neurons is sufficient for decision making (Hayden and Moreno-Bote, 2018). In this framework, a single pool encodes just one of the available options, the one that is under the focus of attention. Previously attended options produce a background activity against which the current option is compared to, and other options fall outside the representation of the neural network (Hayden and Moreno-Bote, 2018; Krajbich et al., 2010; Lim et al., 2011; Redish, 2016; Rich and Wallis, 2016). Thus, comparison and selection between options occurs through a temporal contrast, rather than through mutual inhibition between simultaneously encoded options. This model can be readily extrapolated to multiple many options, with the only dilemma of dividing time or precision into few or many options (like in Fig. 1), thus addressing the associated breadth-depth tradeoffs. The debate of the one-pool versus several-pools models remains open (Hayden and Moreno-Bote, 2018; Ballesta and Padoa-Schioppa, 2019), but electrophysiology experiments with many options should be able to arbitrate between the two hypotheses under the new computational constraints that we have identified here.

3.6 Appendix

Comments and mathematical proofs supporting claims in the main text.

3.6.1 Expected square distance between two random i.i.d. points

The expected square distance between two independently and identically distributed random points $x, y \sim p(\cdot)$ is, by the law of the unconscious statistician,

$$\begin{aligned}\mathbb{E}[d^2(x,y)] &= \int (x-y)^2 p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int x^2 p(x)\,\mathrm{d}x + \int y^2 p(y)\,\mathrm{d}y - 2\int x\,p(x)\,\mathrm{d}x \int y\,p(y)\,\mathrm{d}y \\ &= 2\mathbb{E}[x^2] - 2\mathbb{E}[x]^2 \\ &= 2\mathrm{Var}[x]\end{aligned}$$

3.6.2 Posterior mean of drift is a monotonously increasing function of accumulated evidence

Here we prove that the posterior mean of the drift, which is a random variable with probability distribution defined by Bayes' rule, Eq. (3.3), is a monotonously increasing function of evidence x. We have seen that the expected value of a drift μ given the accumulated evidence x, for any option (and thus here dropping indices) is given by

$$\hat{\mu}(x,t,\sigma,\theta) = \frac{\int \mathrm{d}\mu\,\mu\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)}{p(x|t,\sigma,\theta)}, \tag{3.15}$$

where $p_\theta(\mu)$ is the prior probability of the drifts, with hyperparameters θ and $p(x|t,\sigma,\theta)$ is the marginalized probability distribution of the evidence. To know if $\hat{\mu}(x,t,\sigma,\theta)$ is an increasing function of x, we simply derive, and we expect the derivative to be always positive,

$$\begin{aligned}\frac{\mathrm{d}\hat{\mu}}{\mathrm{d}x} =& \frac{p(x|t,\sigma,\theta)\int \mathrm{d}\mu\,\mu\left(-\frac{x-\mu t}{\sigma^2 t}\right)\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)}{p(x|t,\sigma,\theta)^2} \\ & - \frac{\int \mathrm{d}\mu\,\mu\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)\int \mathrm{d}\mu\left(-\frac{x-\mu t}{\sigma^2 t}\right)\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)}{p(x|t,\sigma,\theta)^2} > 0 \\ \iff & \frac{1}{\sigma^2}\left(\mathbb{E}\left[\mu^2|x,t,\sigma,\theta\right] - \mathbb{E}\left[\mu|x,t,\sigma,\theta\right]^2\right) = \frac{\mathrm{Var}\left[p(\mu|x,t,\sigma,\theta)\right]}{\sigma^2} > 0,\end{aligned}$$

where

$$\mathbb{E}\left[\mu^n|x,t,\sigma,\theta\right] = \frac{\int \mathrm{d}\mu\,\mu^n\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)}{\int \mathrm{d}\mu\,\mathcal{N}(x|\mu t,\sigma^2 t)p_\theta(\mu)}.$$

Since the variance is the expected value of a positive quantity, then we conclude that the expected value of the drift is a monotonously increasing function of the observed accumulated evidence x for any prior.

3.6.3 Expected value of the drift in the small capacity limit

Here we show that in the small capacity limit, the utility in Eq. (3.9) can be written as in Eq. (3.11) for any regular prior distribution. Our strategy is to study the limiting behaviors of the cumulative density function (described below in Sec. 3.6.4) and the posterior mean of the drift (detailed in this section) that appear in Eq. (3.9) as $C = \frac{\sigma_0^2}{\sigma^2}T$ goes to zero.

From Bayes's rule, Eq. (3.3), the posterior mean of the drift is given by

$$\hat{\mu}(x, t, \sigma, \theta) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \frac{\int \mathrm{d}\mu\, \mu \, \exp\left(-\frac{1}{2\sigma^2 t}(\mu t - x)^2\right) p_\theta(\mu|\theta)}{p(x|t, \sigma, \theta)}. \tag{3.16}$$

Let us focus on the numerator, which we will interpret as the expectation value of $\mu \, \exp\left(-\frac{1}{2\sigma^2 t}(\mu t - x)^2\right)$ with respect to the prior. We assume the prior to be such that this expectation is finite for all x and that all its moments are finite (e.g, Gaussian and uniform distributions). We define $z \equiv z(x) \equiv \frac{1}{\sqrt{\sigma^2 t}}(x - \mu_0 t)$ and $\mu_s \equiv \frac{1}{\sqrt{\sigma^2 t}}(\mu t - \mu_0 t)$, and by adding and subtracting $\mu_0 t$ at the exponent, we can write the numerator in the above equation as

$$\begin{aligned}
\mathbb{E}_\theta\left[\mu \, \exp\left(-\frac{1}{2\sigma^2 t}(\mu t - x)^2\right)\right] &= \mathbb{E}_\theta\left[\mu \, \exp\left(-\frac{1}{2\sigma^2 t}(\mu t - \mu_0 t + \mu_0 t - x)^2\right)\right] \\
&= \exp\left(-\frac{z^2}{2}\right)\mathbb{E}_\theta\left[\mu \, \exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right)\right] \\
&= \exp\left(-\frac{z^2}{2}\right)\Bigg\{\mathbb{E}_\theta\left[\mu_0 \, \exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right)\right] \\
&\quad + \mathbb{E}_\theta\left[\sqrt{\frac{\sigma^2}{t}}\mu_s \, \exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right)\right]\Bigg\}.
\end{aligned}$$

Next, we note that the exponential in the expectations is the generating function of the Hermite polynomials, and thus

$$\exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right) = \sum_{n=0}^{\infty} \mathrm{He}_n(z)\frac{\mu_s^n}{n!}.$$

By replacing the exponential with the infinite series in the above expectation, Eq.

(3.16), we obtain

$$
\begin{aligned}
&p(x|t,\sigma,\theta)\hat{\mu}(x,t,\sigma,\theta) = \\
&\quad =\frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi\sigma^2 t}}\left\{\mathbb{E}_\theta\left[\mu_0\sum_{n=0}^{\infty}\mathrm{He}_n(z)\frac{\mu_s^n}{n!}\right]+\mathbb{E}_\theta\left[\sqrt{\frac{\sigma^2}{t}}\mu_s\sum_{n=0}^{\infty}\mathrm{He}_n(z)\frac{\mu_s^n}{n!}\right]\right\} \\
&\quad =\frac{\mathcal{N}(z|0,1)}{\sqrt{\sigma^2 t}}\left\{\sum_{n=0}^{\infty}\frac{1}{n!}\mathbb{E}_\theta\left[\mu_s^n\right]\mu_0\mathrm{He}_n(z)+\sum_{n=0}^{\infty}\frac{1}{n!}\mathbb{E}_\theta\left[\mu_s^{n+1}\right]\sqrt{\frac{\sigma^2}{t}}\mathrm{He}_n(z)\right\} \\
&\quad =\frac{\mathcal{N}(z|0,1)}{\sqrt{\sigma^2 t}}\left\{\sum_{n=0}^{\infty}\frac{1}{n!}\mathbb{E}_\theta\left[\frac{(\mu-\mu_0)^n}{\sqrt{\sigma^2/t}^n}\right]\left(\mu_0\mathrm{He}_n(z)+\sqrt{\frac{\sigma^2}{t}}n\mathrm{He}_{n-1}(z)\right)\right\} \\
&\quad =\frac{\mathcal{N}(z|0,1)}{\sqrt{\sigma^2 t}}\left\{\sum_{n=0}^{\infty}\frac{1}{n!}\sqrt{\frac{C}{M}}^{n-1}\mathbb{E}_\theta\left[\frac{(\mu-\mu_0)^n}{\sigma_0^n}\right]\right. \\
&\qquad\qquad \left.\times\left(\sqrt{\frac{C}{M}}\mu_0\mathrm{He}_n(z)+\sigma_0 n\mathrm{He}_{n-1}(z)\right)\right\},
\end{aligned}
$$

where we have used that all the moments of the prior are finite and the sum is well defined. Note that to obtain the third line we have shifted the second index $n+1\rightarrow n$ and used that the term $n\mathrm{He}_{n-1}(z)$ is zero for $n=0$.

We now insert the above series into the expression of utility in Eq. (3.9) to obtain

$$
\begin{aligned}
\hat{U}(t,\sigma,\theta) &= \int \mathrm{d}x\ \frac{\mathrm{d}}{\mathrm{d}x}\left\{\left[F_x(x|t,\sigma,\theta)\right]^M\right\}\hat{\mu}(x,t,\sigma,\theta) \\
&= \int \mathrm{d}x\ M\left[F_x(x|t,\sigma,\theta)\right]^{M-1}\frac{\mathcal{N}(z|0,1)}{\sqrt{\sigma^2 t}} \\
&\quad\times\left\{\sum_{n=0}^{\infty}\frac{1}{n!}\sqrt{\frac{C}{M}}^{n-1}\mathbb{E}_\theta\left[\frac{(\mu-\mu_0)^n}{\sigma_0^n}\right]\left(\sqrt{\frac{C}{M}}\mu_0\mathrm{He}_n(z)+\sigma_0 n\mathrm{He}_{n-1}(z)\right)\right\} \\
&= \int \mathrm{d}z\ M\left[F_z(z|t,\sigma,\theta)\right]^{M-1}\mathcal{N}(z|0,1)\left[\mu_0+\sqrt{\frac{C}{M}}\sigma_0 z\right]+\mathcal{O}(C),
\end{aligned}
$$

where in the second line it is implicit that z depends on x, and in the last line we have made a linear transformation of variables from x to $z=z(x)$. We also note that as the integral in the last line only involves polynomials in z that are weighted by the standard normal (and by a cumulative, which is bounded to be in the range $[0,1]$), their integrals are finite, and thus we can truncate the series at the first leading order, which is order $\sqrt{C}$. It remains to see whether the cumulative density function $F_z(z|t,\sigma,\theta)$ contributes order $\sqrt{C}$ or larger, and we show below in Sec. (3.6.4) that the former is actually true, such that $F_z(z|t,\sigma,\theta) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right] + \mathcal{O}(C)$. With all this, we can approximate the utility up to order $\sqrt{C}$ as

$$\hat{U}(C, M, \mu_0) = \mu_0 + \sigma_0\sqrt{C}\sqrt{M}\int_{-\infty}^{\infty} \mathrm{d}z \left[\frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right)\right]^{M-1} \mathcal{N}(z|0,1)\, z + \mathcal{O}(C), \tag{3.17}$$

which is identical to Eq. (3.11) in main manuscript.

3.6.4 Distribution of evidence at small capacity limit

Here, we find an approximation to the marginalized probability distribution of the evidence at small capacity. From Bayes' rule and the law of the unconscious statistician,

$$p(x|t,\sigma,\theta) = \int \mathrm{d}\mu\, \mathcal{N}(x|\mu t, \sigma^2 t) p_\theta(\mu|\theta) = \mathbb{E}_\theta\left[\mathcal{N}(x|\mu t, \sigma^2 t)\right].$$

To compute this expectation, we follow the same procedure as in Sec. 3.6.3. We define $z \equiv \frac{1}{\sqrt{\sigma^2 t}}(x - \mu_0 t)$ and $\mu_s \equiv \frac{1}{\sqrt{\sigma^2 t}}(\mu t - \mu_0 t)$ and add and subtract $\mu_0 t$ at the exponent, to obtain

$$\begin{aligned} p(x|t,\sigma,\theta) &= \mathbb{E}_\theta\left[\exp\left(-\frac{1}{2\sigma^2 t}(x - \mu t)^2\right)\right] \\ &= \mathbb{E}_\theta\left[\exp\left(-\frac{1}{2\sigma^2 t}(x - \mu_0 t + \mu_0 t - \mu t)^2\right)\right] \\ &= \exp\left(-\frac{z^2}{2}\right)\mathbb{E}_\theta\left[\exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right)\right]. \end{aligned}$$

Next, we again identify the exponential generating function of the Hermite polynomials,

$$\exp\left(z\mu_s - \frac{1}{2}\mu_s^2\right) = \sum_{n=0}^{\infty} \mathrm{He}_n(z)\frac{\mu_s^n}{n!},$$

and thus we obtain a series for the probability distribution of the evidence,

$$\begin{aligned}
p(x|t,\sigma,\theta) =& \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{z^2}{2}\right) \mathbb{E}_\theta \left[\sum_{n=0}^{\infty} \mathrm{He}_n(z) \frac{\mu_s^n}{n!}\right] \\
=& \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi\sigma^2 t}} \sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E}_\theta \left[\mu_s^n\right] \mathrm{He}_n(z) \\
=& \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi\sigma^2 t}} \sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E}_\theta \left[\frac{(\mu-\mu_0)^n}{\sqrt{\sigma^2/t}^n}\right] \mathrm{He}_n(z) \\
=& \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi\sigma^2 t}} \sum_{n=0}^{\infty} \frac{1}{n!} \sqrt{\frac{C}{M}}^n \mathbb{E}_\theta \left[\frac{(\mu-\mu_0)^n}{\sigma_0^n}\right] \mathrm{He}_n(z).
\end{aligned}$$

We see that the leading order the distribution of the evidence is a normal distribution, while the order $\sqrt{C}$ is zero. Therefore, its cumulative in the variable $z = z(x)$ is, exactly, up to order $\sqrt{C}$, $F_z(z|t,\sigma,\theta) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] + \mathcal{O}(C)$. This expression has been used in Sec. (3.6.3).

3.6.5 Asymptotic limit of relevant integral

In this subsection we want to obtain the asymptotic limit, $M \to \infty$, of the integral

$$I(M) = \int_{-\infty}^{\infty} \mathrm{d}y \, y \, \frac{\mathrm{d}}{\mathrm{d}y} \Phi^M(y),$$

appearing in Eqs. (3.11) and (3.12), where $\Phi(y)$ is the normal cumulative distribution function,

$$\Phi(y) = \left(\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{y}{\sqrt{2}}\right)\right).$$

Using Extreme Value Theory (De Haan and Ferreira, 2007), it can be shown that this cumulative distribution function $\Phi(y)$ belongs to the Gumbel class of the generalized extreme value distributions,

$$\lim_{M\to\infty} \Phi^M(a_M y + b_M) = G(y),$$

where $G(y) = \exp(-\exp(-y))$ and

$$b_M = (2\log(M) - \log(\log(M)) - \log(4\pi))^{1/2} \quad \text{and} \quad a_M = 1/b_M.$$

Using this result, then our integral develops quite easily,

$$\begin{aligned} I(M) &\to \int_{-\infty}^{\infty} \mathrm{d}y\, y\, \frac{\mathrm{d}}{\mathrm{d}y} G\left(\frac{y-b_M}{a_M}\right) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y \left(\frac{y}{b_M} + b_M\right) \frac{\mathrm{d}}{\mathrm{d}y} G(y) \\ &= \frac{1}{b_M} \int_{-\infty}^{\infty} \mathrm{d}y\, y \exp(-y) \exp(-\exp(-y)) + b_M \int_{-\infty}^{\infty} \mathrm{d}y\, \frac{\mathrm{d}}{\mathrm{d}y} G(y) \\ I(M \to \infty) &= \frac{\gamma}{b_M} + b_M, \end{aligned}$$

where $\gamma \approx 0.577$ is Euler's constant.

3.6.6 Expected utility for uniform prior

For this choice of prior, drifts are all drawn independently and identically from a uniform probability distribution between zero and one. That is,

$$p(\mu_i) = \Theta(\mu_i)\Theta(1-\mu_i),$$

where $\Theta(x)$ is the Heaviside step function. We can substitute this prior into eq. (3.3) to obtain the posterior probability distribution for the drifts,

$$p(\mu_i|x_i, \sigma, t_i, \theta) = \begin{cases} \frac{\mathcal{N}\left(\mu_i|\frac{x_i}{t_i}, \frac{\sigma^2}{t_i}\right)}{\int_0^1 \mathcal{N}\left(\mu_i|\frac{x_i}{t_i}, \frac{\sigma^2}{t_i}\right)\mathrm{d}\mu_i} & \mu_i \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

This will produce an expectation value for each drift,

$$\hat{\mu}_i(x_i, t_i, \sigma) \equiv \mathbb{E}\left[\mu_i|x_i, t_i, \sigma\right] = \frac{x_i}{t_i} + \frac{\sigma}{\sqrt{2\pi t_i}} \frac{\exp\left(-\frac{x_i^2}{2\sigma^2 t_i}\right) - \exp\left(-\frac{(x_i-t_i)^2}{2\sigma^2 t_i}\right)}{\frac{1}{2}\left[\operatorname{erf}\left(\frac{x_i}{\sqrt{2\sigma^2 t_i}}\right) - \operatorname{erf}\left(\frac{x_i-t_i}{\sqrt{2\sigma^2 t_i}}\right)\right]}, \tag{3.18}$$

where the denominator is related to the probability distribution of the evidence x_i, which we can find by marginalizing over drifts,

$$\begin{aligned} p(x_i|t_i, \sigma) &= \int_0^1 \mathrm{d}\mu \frac{1}{\sqrt{2\pi\sigma^2 t_i}} \exp\left(-\frac{1}{2\sigma^2 t_i}(x_i - \mu_i t_i)^2\right) \\ &= \frac{1}{2t_i}\left[\operatorname{erf}\left(\frac{x_i}{\sqrt{2\sigma^2 t_i}}\right) - \operatorname{erf}\left(\frac{x_i - t_i}{\sqrt{2\sigma^2 t_i}}\right)\right]. \end{aligned} \tag{3.19}$$

We will use from now on the assumption of even time allocation, $t_i = t = \frac{T}{M}$ for all i. The cumulative probability distribution for the evidence in Eq. (3.19) is, integrating by parts,

$$\begin{aligned}
F(x|t,\sigma) &= \int_{-\infty}^{x} p(x'|t,\sigma)\,\mathrm{d}x' \\
&= \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-t}{\sqrt{2\sigma^2 t}}\right)\right) + \frac{x}{2t}\left(\operatorname{erf}\left(\frac{x}{\sqrt{2\sigma^2 t}}\right) - \operatorname{erf}\left(\frac{x-t}{\sqrt{2\sigma^2 t}}\right)\right) \\
&+ \sqrt{\frac{\sigma^2}{2\pi t}}\left[\exp\!\left(-\frac{x^2}{2\sigma^2 t}\right) - \exp\!\left(-\frac{(x-t)^2}{2\sigma^2 t}\right)\right] \\
&= \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-t}{\sqrt{2\sigma^2 t}}\right)\right] + tp(x|t,\sigma)\hat{\mu}(x,t,\sigma),
\end{aligned}$$

where in the last equality we have rewritten the solution in a convenient form. Hence, the product of the expected value with the probability density can be rewritten in terms of the cumulative function, from the previous equation,

$$\hat{\mu}(x,t,\sigma)p(x|t,\sigma) = \frac{1}{t}F(x|t,\sigma) - \frac{1}{2t}\left[1+\operatorname{erf}\left(\frac{x-t}{\sqrt{2\sigma^2 t}}\right)\right],$$

and using eq. (3.9) we get the expression for the utility,

$$\hat{U}(M,t,\sigma) = \frac{M}{t}\int_{-\infty}^{\infty}\mathrm{d}x\,[F(x|t,\sigma)]^{M-1}\left\{F(x|t,\sigma) - \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-t}{\sqrt{2\sigma^2 t}}\right)\right]\right\}. \tag{3.20}$$

3.6.7 Expected utility for Gaussian bimodal prior

The expected utility for the bimodal Gaussian prior with modes μ_1 and μ_2, each with a variance σ_0^2, is quite similar to the unimodal, Eq. (3.10), and follows the straightforward application of Eq. (3.9). The probability distribution of the evidence marginalized over drifts is $p(x|t,\sigma,\theta) = \frac{1}{2}\mathcal{N}(x|\mu_1 t, \sigma^2 t + \sigma_0^2 t^2) + \frac{1}{2}\mathcal{N}(x|\mu_2 t, \sigma^2 t + \sigma_0^2 t^2)$. Therefore the cumulative is

$$F(x|t,\sigma,\theta) = \frac{1}{2}\Phi(x|\mu_1 t, \sigma^2 t + \sigma_0^2 t^2) + \frac{1}{2}\Phi(x|\mu_2 t, \sigma^2 t + \sigma_0^2 t^2),$$

where $\Phi(x|\mu_m, \sigma_m^2)$ is the normal cumulative distribution for one mode. However, the expected value of the drift is a bit more involved, since the posterior distribution over drifts takes a different form,

$$\begin{aligned}
p(\mu|x,t,\sigma,\theta) =& \frac{\sigma_t}{\sqrt{2\pi\sigma^2 t\sigma_0^2}p(x|t,\sigma,\theta)} \\
&\times \sum_i \frac{1}{2}\mathcal{N}(\mu|\hat{\mu}_i, \sigma_t^2)\exp\left(-\frac{1}{2(\sigma_0^2 t^2 + \sigma^2 t)}\left(\mu_i t - x\right)^2\right),
\end{aligned}$$

where $1/\sigma_t^2 = t/\sigma^2 + 1/\sigma_0^2$ and

$$\hat{\mu}_i = \frac{\sigma_t^2}{\sigma_0^2}\mu_i + \frac{\sigma_t^2}{\sigma^2}x.$$

Consequently, the expected value will be

$$\hat{\mu}(x, t, \sigma, \theta) = \frac{1}{\sqrt{2\pi(\sigma_0^2 t^2 + \sigma^2 t)}} \frac{1}{p(x|t, \sigma, \theta)} \sum_i \frac{\hat{\mu}_i}{2} \exp\left(-\frac{1}{2(\sigma_0^2 t^2 + \sigma^2 t)} (\mu_i t - x)^2\right).$$

Then, expected utility is

$$\hat{U}(M, t, \sigma, \theta) = M \int_{-\infty}^{\infty} \mathrm{d}x\, F(x|t, \sigma, \theta)^{M-1} \sum_i \frac{\hat{\mu}_i}{2} \mathcal{N}\left(x|\mu_i t, \sigma_0^2 t^2 + \sigma^2 t\right). \tag{3.21}$$

This expression is numerically integrated and used in Fig. 3.4.

3.6.8 Stochastic gradient ascent method for Gaussian prior

To maximize utility, Eq. (3.14) in main manuscript, under the time constraint, we can make use of unconstrained optimization through Lagrangian multipliers. We construct the Lagrangian given by

$$L(\mathbf{t}, \theta, \lambda) = \hat{U}(\mathbf{t}, \sigma, \theta) + \lambda h(\mathbf{t}) + \boldsymbol{\mu} \cdot \mathbf{g}(\mathbf{t}), \tag{3.22}$$

where $h(\mathbf{t}) = \sum_{i=1}^{M} t_i - T$ is the equality constraint that defines the hyperplane and $0 \leq g_i(\mathbf{t}) = t_i$ is the inequality constraint forcing all times i to be non-negative and thus defining the simplex. The quantities λ and $\boldsymbol{\mu}$ are the Lagrangian multipliers. In other words, maximizing utility, Eq. (3.14), subject to the initial constraints can be done by optimizing the Lagrangian, Eq. (3.22), with respect to $\mathbf{t}$, λ and $\boldsymbol{\mu}$ subject to Karush-Kuhn-Tucker conditions (Bishop, 2006)

$$g_i(\mathbf{t}) \geq 0, \quad \text{for all } i \tag{3.23a}$$

$$\mu_j \geq 0, \quad \text{for all } j \tag{3.23b}$$

$$\boldsymbol{\mu} \cdot \mathbf{g}(\mathbf{t}) = 0 \tag{3.23c}$$

We notice that the first two conditions imply that the third can be rewritten as $\mu_i t_i = 0$ for all i. By optimizing the Lagrangian, Eq. (3.22), we obtain the following system of equations

$$\nabla_{\mathbf{t}} \hat{U}(\mathbf{t}^*, \sigma, \theta) + \lambda^* \mathbf{1} + \boldsymbol{\mu}^* = \mathbf{0}, \tag{3.24}$$

where $\mathbf{t}^*, \lambda^*, \boldsymbol{\mu}^*$ denote the critical points of the Lagrangian. We note that the symmetrical point in M active dimensions, denoted by $\mathbf{t}_M^e$, where $t_{M,i}^e = T/M$

for $i = 1, ..., M$, is a critical point of the Lagrangian. This is because the partial derivatives with respect to the utility have to be equal at $\mathbf{t}_M^e$, and since this point lies in the interior of the $(M-1)$-simplex, the $\boldsymbol{\mu}_i^e = 0$ for $i = 1, \ldots, M$. Therefore $\mathbf{t}_M^e$ complies with Eq. (3.24) and is indeed a critical point.

Next, we detail the gradient ascent method used to obtain Fig. 3.5**e**. As explained above, we want to optimize utility, Eq. (3.14), subject to a set of equality, Eq. (3.1), and inequality constraints, $t_i \geq 0$, as described in section "Even allocation is optimal" of the Results. As all our constraints are linear, we can make use of the gradient projection method (Fletcher, 2013). In this case, we want to obtain the gradient of utility in Eq. (3.10) and project it in the $(M-1)$-simplex such that the capacity constraint in Eq. (3.1) is satisfied. Due to the linear capacity equality constraint, this projection is simply given by the linear operator

$$\Pi = \mathbf{Id}_{M\times M} - \frac{1}{M}\mathbf{1}_{M\times M}$$

where $\mathbf{Id}_{M\times M}$ is the $M \times M$ identity matrix and $\mathbf{1}_{M\times M}$ is an $M \times M$ matrix full of ones. Therefore, we can maximize utility by updating $\mathbf{t}^{(k)}$ appropriately,

$$\mathbf{t}^{(k+1)} = \mathbf{t}^{(k)} + \eta\Pi\left(\nabla_{\mathbf{t}}\,\hat{U}(\mathbf{t}, \sigma, \theta)\right), \tag{3.25}$$

where $\eta = 10^{-1}T$ is the default step size, k is the iteration number, and θ corresponds to the parameters of the Gaussian prior. The utility for an arbitrary time allocation $\mathbf{t}$ for the Gaussian prior case is, using Eq. (3.14),

$$\hat{\mathbf{U}}(\mathbf{t}, \sigma, \theta) = \mu_0 + \sigma_0 \sum_{i=1}^{N} \int_{-\infty}^{\infty} \mathrm{d}y\; y\, \frac{\exp\left(-\frac{y^2}{2\sigma_i^2}\right)}{\sqrt{2\pi\sigma_i^2}} \prod_{j\neq i} \left\{\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{y}{\sqrt{2}\sigma_j}\right)\right]\right\}, \tag{3.26}$$

where $\sigma_i^2 = \frac{\sigma_0^2 t_i}{\sigma_0^2 t_i + \sigma^2}$. We can therefore compute the derivative of the previous equation with respect to all components t_i and numerically integrate the expression that results.

In addition to the linear capacity constraint, we have to enforce the inequality constraints as well, i.e. $t_i \geq 0$, which we do by utilizing an active set of constraints. To implement it, we start in a relatively high-dimensional $(M-1)$-simplex, choosing M to be $2M^*$, where M^* is the optimal number of accumulators to sample in the even sampling case (which is estimated before through exploration, see main text). If and whenever any of the components of $\mathbf{t}^{k+1}$ derived from Eq. (3.25) is approaching a border ($t_i^{(k+1)} \approx \tau$ for some i and small τ), the step size decreases until the component effectively reaches zero. In such a case, this dimension is added to the active constraints set (we inactivate the dimension), thus downgrading the simplex to a lower dimension. In this way, our algorithm only reduces the initial dimension of the simplex and never extends it. To initially activate the $2M^*$ dimensions, for any random initial condition $\mathbf{t}_0$, we make sure that all the components are greater than our threshold $t_{0,i} > \tau$ for all $i = 1, ..., 2M^*$.

Finally, in order to avoid potentially getting trapped in local maxima, we add noise at every iteration as follows. At every step k of Eq. (3.25), and with probability $\epsilon = 0.1$, we push the $t_i^{(k)}$ of a randomly chosen dimension i by a magnitude $\delta = 10^{-3}T$ and pull the $t_j^{(k)}$ of another random dimension j in the opposite direction with the same amount in order to stay in the appropriate simplex.

Chapter 4

OPTIMALITY WITH INTRINSIC MOTIVATION

The following chapter is based on the pre-print manuscript currently in https://arxiv.org/abs/2205.10316*. See (Ramírez-Ruiz et al., 2022) for bibliographic details. This version is currently under consideration. I declare to be the main author of this work.*

Abstract

Most theories of behavior posit that agents tend to maximize some form of reward or utility. However, animals very often move with curiosity and seem to be motivated in a reward-free manner. Here we abandon the idea of reward maximization, and propose that the sole goal of intelligent behavior is maximizing occupancy of future paths of actions and states, a principle that we call path occupancy maximization. According to this view, rewards are the means to occupy path space, not the goal per se; goal-directedness simply emerges as rational ways of searching for resources so that movement, understood amply, never ends. We find that action-state path entropy is the only measure consistent with additivity and other intuitive properties of expected future action-state path occupancy. We provide analytical expressions that relate the optimal policy and state-value function, and prove convergence of our value iteration algorithm. Using discrete and continuous state tasks, we show that complex behaviors such as 'dancing', hide-and-seek and a basic form of altruistic behavior naturally result from the intrinsic motivation to occupy path space. All in all, we present a theory of behavior that generates goal-directedness in the absence of reward maximization.

4.1 Introduction

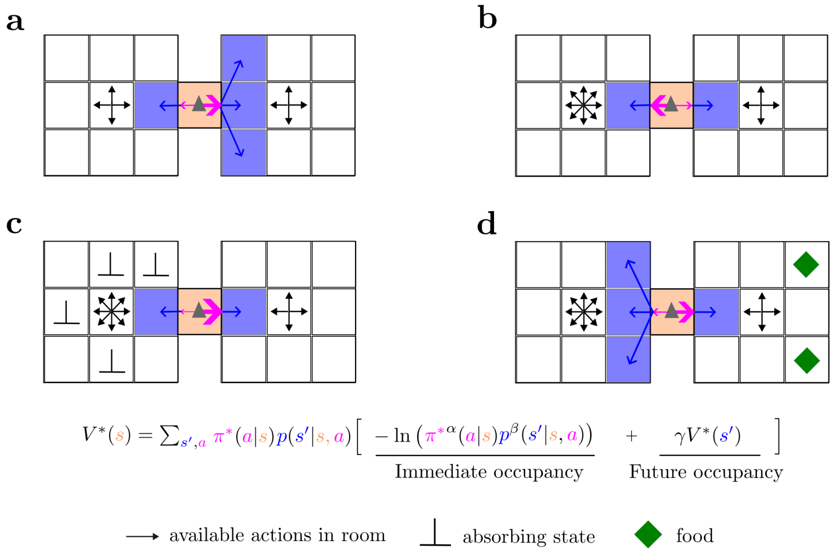
Natural agents are endowed with a natural tendency to move, explore and interact with their environment (Ryan and Deci, 2000; Oudeyer et al., 2007). For instance, human newborns unintentionally move their body parts (Adolph and Berger, 2007), and 7 to 12-months infants spontaneously babble vocally (MacNeilage and Davis, 2000) and with their hands (Petitto and Marentette, 1991). Exploration and curiosity are major drives for learning and discovery through information-seeking (Dietrich, 2004; Kidd and Hayden, 2015; Gottlieb et al., 2013). These behaviors seem to elude a simple explanation in terms of external reward maximization. However, intrinsic motivation pushes agents to visit new states by performing novel courses of action, which helps learning and the discovery of even larger rewards in the long run (Gittins et al., 2011; Averbeck, 2015). Therefore, it has been argued that exploration and curiosity could have arisen as a consequence of seeking external reward maximization by endowing agents with the necessary inductive biases to learn in complex and ever-changing natural environments (Doll et al., 2012; Wang and Hayden, 2021).

While most theories of rational behavior posit that agents are reward maximizers (von Neumann and Morgenstern, 1953; Sutton et al., 1998; Kahneman and Tversky, 2013), very few of us would agree that the sole goal of living agents is maximizing money gains or food intake. Indeed, expressing excessive emphasis on those goals is a sign of psychological disorders (Rash et al., 2016; Ágh et al., 2016). Further, setting a reward function by design as the goal of artificial agents is more often than not arbitrary (Sutton et al., 1998; McNamara and Houston, 1986; Klyubin et al., 2005; Lehman and Stanley, 2011), leading to the recurrent problem faced by theories of reward maximization of defining what rewards are (Singh et al., 2009; Zhang et al., 2021b; Schmidhuber, 1991b; Hadfield-Menell et al., 2017; Eysenbach et al., 2018). In some cases, like in artificial games, rewards can be unambiguously defined, such as number of collected points or wins (Schrittwieser et al., 2020). However, in most situations defining rewards is task-dependent, non-trivial and problematic. For instance, a vacuum cleaner robot could be designed to either maximize the weight or volume of dust collected, energy efficiency, or a weighted combination of them (Asafa et al., 2018). In more complex cases, companies can aim at maximizing profit, but without a suitable innovation policy profit maximization can be self-defeating (Kline and Rosenberg, 2010).

Here, we abandon the idea that the goal is maximizing external reward and that movement over space is a means to achieve this goal. Instead, we adopt the opposite view, inspired by the nature of our intrinsic drives: we propose that the objective *is* to maximally occupy action-state path space, understood in a broad sense, in the long term. We call this principle path occupancy maximization (POM), which posits that the goal of agents is to generate all sort of behaviors and occupy, on average, as much space (action-state paths) as possible in the future. According to this principle, external rewards serve to generate the work necessary to occupy action-state space, they are not the goals per se. The usual exploration–exploitation tradeoff (Wilson et al., 2021) therefore disappears: agents that seek to occupy space "solve" this issue naturally because they care about rewards only as means to an end. Furthermore, in this sense, surviving is only preferred because it is needed to keep visiting action-state space. Our theory provides a rational account of exploratory and curiosity-driven behavior where the problem of defining an external reward function vanishes, and captures the variability of behavior (Moreno-Bote et al., 2011; Recanatesi et al., 2022; Corver et al., 2021; Dagenais et al., 2021; Mochol et al., 2021; Cazettes et al., 2021) by taking it as a principle.

We build over an extensive literature on entropy-regularized reinforcement learning (RL) (Todorov, 2009; Ziebart, 2010; Haarnoja et al., 2018; Schulman et al., 2017; Neu et al., 2017; Hausman et al., 2018; Tishby and Polani, 2011; Nachum et al., 2017; Galashov et al., 2019). While that literature emphasizes the regularization benefits of entropy for learning, external rewards still serve as the major drive of behavior. Intrinsic motivation approaches where reward does not play any role are closer to ours. One type of reward-free approaches promotes exploration through information-seeking objectives, such as minimizing surprise by refining predictions (Burda et al., 2018; Achiam and Sastry, 2017; Fountas et al., 2020; Burda et al., 2019; Pathak et al., 2017; Hafner et al., 2020) or novelty seeking (Bellemare et al., 2016; Tang et al., 2017; Aubret et al., 2022). One central prediction of these approaches is that exploration, and thus behavioral variability, ceases after learning. Our POM principle is different in that the objective is determined independently from the agents' knowledge of the environment, and thus movement through the occupation of path space never ends, even if there is nothing to learn. Another type of intrinsic motivation approaches that are close to ours are pure entropic objectives, but they concentrate instead on the coverage problem, maximizing the stationary state entropy (Hazan et al., 2019; Liu and Abbeel, 2021; Mutti et al., 2021; Seo et al., 2021; Zhang et al., 2021a; Amin et al., 2021), and they do so typically to generate better policies in the exploitation phase when a well-defined task is to be solved. Finally, there is a class of reward-free objectives known as empowerment that focus on generating policies under which state transitions are predictable (Klyubin et al., 2005; Jung et al., 2011; Still and Precup, 2012; Mohamed and Jimenez Rezende, 2015), which is different from occupying path space.

In this work, we model an agent interacting with the environment as a Markov decision process (MDP) where the intrinsic, immediate reward is the occupancy of the next action-state visited, which is largest when performing an uncommon action and visiting a rare state –there are no external rewards that drive the agent. We assume that the agent maximizes the occupancy of future action-state paths. We



$$V^*(s) = \textstyle\sum_{s',a} \pi^*(a|s)p(s'|s,a)\Big[\underbrace{-\ln\left(\pi^{*\alpha}(a|s)p^{\beta}(s'|s,a)\right)}_{\text{Immediate occupancy}} + \underbrace{\gamma V^*(s')}_{\text{Future occupancy}} \Big]$$

Figure 4.1: Entropy-seeking agents (H agents) achieve path occupancy maximization (POM). An H agent (grey triangle) in the middle of two rooms has the choice between going left or right. Black arrows indicate available actions, blue arrows indicate random transitions after choosing moving right or moving left actions, and pink arrow width indicates the probabilities of those actions. See text for details.

show that action-state path entropy is the only measure of occupancy consistent with additivity per time step, positivity and smoothness. Due to the additivity property, the value of being in a state, defined as the expected future time-discounted action-sate path entropy, can be written in the form of a Bellman equation. We show that the Bellman equation has a unique solution that can be found with an iterative map. In four simulated experiments we show that the sole goal of maximizing future action-state path entropy generates complex behaviors that, to the human eye, look genuinely goal-directed and playful, such as hide-and-seek in a prey-predator problem, dancing of a cartpole and a basic form of altruism in an agent-and-pet example.

4.2 Path occupancy maximization principle

4.2.1 Entropy measure of path space occupancy

We model an agent as a finite action-state MDP in discrete time. The policy π describes the probability $\pi(a|s)$ of performing action a given that the agent is at state s at some time step, and $p(s'|s,a)$ is the transition probability from s to a successor state s' in the next time step given that action a is performed. Starting at

$t = 0$ in state s_0, an agent performing a sequence of actions and experiencing state transitions $\tau \equiv (s_0, a_0, s_1, ..., a_t, s_{t+1}, ...)$ gets a return defined as

$$R(\tau) = -\sum_{t=0}^{\infty} \gamma^t \ln\left(\pi^{\alpha}(a_t|s_t) p^{\beta}(s_{t+1}|s_t, a_t)\right) \tag{4.1}$$

with action and state weights $\alpha > 0$ and $\beta \geq 0$, respectively, and discount factor $0 < \gamma < 1$. A larger return is obtained when, starting in s_t, a low-probability action a_t is performed and followed by a low-probability transition to a state s_{t+1}. Therefore, maximizing the return in Eq. (4.1) favors 'visiting' action-states (a_t, s_{t+1}) with a low transition probability. From s_{t+1}, another low-probability action-state transition is preferred and so on, such that low-probability trajectories τ are more rewarding than high-probability ones. Thus, the agent is pushed to visit action-states that are rare or 'unoccupied', implementing POM. Due to the freedom to choose action a_t given state s_t and the uncertainty of the resulting next state s_{t+1}, apparent in Eq. (4.1), the term 'action-states' used here is more natural than 'state-actions'.

The agent is assumed to optimize the policy π to maximize the state-value $V_\pi(s)$, defined as the expected return

$$V_\pi(s) \equiv \mathbb{E}_\pi[R(\tau)|s_0 = s] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left(\alpha \mathcal{H}(A|s_t) + \beta \mathcal{H}(S'|s_t, a_t)\right) \Big| s_0 = s\right] \tag{4.2}$$

given the initial condition $s_0 = s$ and following policy π, that is, the expectation is over the $a_t \sim \pi(a_t|s_t)$ and $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$, $t \geq 0$. In the last identity, we have rewritten the expectations of the terms in Eq. (4.1) as a discounted sum of action and successor state conditional entropies $\mathcal{H}(A|s) = -\sum_a \pi(a|s) \ln \pi(a|s)$ and $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a)$, respectively. We stress that this expected return is purely intrinsic, namely, there is no external reward (policy-independent reinforcer) that the agent seeks to maximize.

We call entropy-seeking agent (H agent) the one that optimizes the policy to maximize the state-value in Eq. (4.2). The entropy representation in Eq. (4.2) of the POM principle has several implications. First, H agents prefer regions of state space that lead to a large number of successor states (Fig. 4.1a) or larger number of actions (Fig. 4.1b). Second, death (absorbing) states where only one action-state (i.e., stay) is available forever are naturally avoided by an H agent, as they promise zero future action and state entropy (Fig. 4.1c). Therefore, our framework implicitly incorporates a survival instinct. Finally, regions of state space where there are "rewarding" states that increase the capacity of the agent to visit further action-states (such as filling an energy reservoir) are more frequently visited than others (Fig. 4.1d).

We found that maximizing the discounted action-state path entropy in Eq. (4.2) is the only reasonable way of formalizing the POM principle, as it is the only measure of action-state path occupancy in Markov chains consistent with the following intuitive conditions: if a path τ has probability p, visiting it results in an occupancy gain $C(p)$ that (i) decreases with p and (ii) is first-order differentiable (Supplemental Sec. 4.5.1 for details). Condition (i) implies that visiting a low probability path increases occupancy more than visiting a high probability path, and our agents should tend to occupy 'unoccupied' path space; condition (ii) requires that the measure should be smooth. We also ask that (iii) the occupancy of paths, defined as the expectation of occupancy gains over paths given a policy, is the sum of the expected occupancies of their subpaths (additivity condition). This last condition implies that agents can accumulate occupancy over time by keeping visiting low-probability action-states, but the accumulation should be consistent with the Markov property of the decision process.

4.2.2 Optimal policy and state-value function

The state-value $V_\pi(s)$ in Eq. (4.2) can be recursively written using the values of successor states through the standard Bellman equation

$$\begin{aligned} V_\pi(s) &= \alpha\mathcal{H}(A|s) + \beta\sum_a \pi(a|s)\mathcal{H}(S'|s,a) + \gamma\sum_{a,s'}\pi(a|s)p(s'|s,a)V_\pi(s') \\ &= \sum_{a,s'}\pi(a|s)p(s'|s,a)\left(-\alpha\ln\pi(a|s) - \beta\ln p(s'|s,a) + \gamma V_\pi(s')\right), \quad (4.3) \end{aligned}$$

where the sum is over the available actions a from state s and over the successor states s' given the performed action at state s. The optimal policy π^* that maximizes the state-value is defined as $\pi^* = \arg\max_\pi V_\pi$ and the optimal state-value is

$$V^*(s) = \max_\pi V_\pi(s), \quad (4.4)$$

where the maximization is with respect to the $\{\pi(\cdot|\cdot)\}$ for all actions and states. To obtain the optimal policy, we first determine the critical points of the expected return $V_\pi(s)$ in Eq. (4.3) using Lagrange multipliers (Supplemental Sec. 4.5.2). The optimal state-value $V^*(s)$ is found to obey the non-linear self-consistency set of equations

$$\begin{aligned} V^*(s) &= \alpha\ln Z(s) \quad (4.5) \\ &= \alpha\ln\left[\sum_a \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma\sum_{s'}p(s'|s,a)V^*(s')\right)\right], \quad (4.6) \end{aligned}$$

where $Z(s)$ is the partition function, defined by substitution, and the critical policy satisfies

$$\pi^*(a|s) = \frac{1}{Z(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma\sum_{s'} p(s'|s,a)V^*(s')\right). \quad (4.7)$$

We find that the solution to the non-linear system of Eqs. (4.6) is unique and, moreover, the unique solution is the absolute maximum of the state-values over all policies (Supplemental Sec. 4.5.3).

To determine the actual value function from such non-linear set of equations, we derive an iterative map, a form of value iteration that exactly incorporates the optimal policy at every step. Defining $z_i = \exp(\alpha^{-1}\gamma V(s_i))$, $p_{ijk} = p(s_j|s_i,a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i,a_k)$, Eq. (4.6) can be turned into the iterative map

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)}\right)^{p_{ijk}}\right)^{\gamma} \quad (4.8)$$

for $n \geq 0$ and with initial conditions $z_i^{(0)} > 0$. Here, the matrix with coefficients $w_{ik} \in \{0,1\}$ indicate whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$), and j extends over all states, with the understanding that if a state s_j is not a possible successor from state s_i after performing action a_k then $p_{ijk} = 0$. We find that the infinite series $z_i^{(n)}$ defined in Eq. (4.8) converges to a finite limit $z_i^{(n)} \to z_i^\infty$ regardless of the initial condition in the positive first orthant, and that $V^*(s_i) = \alpha\gamma^{-1}\ln z_i^\infty$ is the optimal state-value function, which solves Eq. (4.6) (Supplemental Sec. 4.5.3). Iterative maps similar to Eq. (4.8) have been studied before (Todorov, 2009, 2006), subsequently shown to have uniqueness (Rubin et al., 2012) and convergence guarantees (Nachum et al., 2017; Leibfried et al., 2019) in the absence of state entropy terms. A summary of results and particular examples can be found in Supplemental Sec. 4.5.4.

We note that in the definition of return in Eq. (4.2) we could replace the absolute action entropy terms $\mathcal{H}(A|s)$ by relative entropies of the form

$$-D_{\mathrm{KL}}(\pi(a|s)||\pi_0(a|s)) = \sum_a \pi(a|s)\ln(\pi_0(a|s)/\pi(a|s)),$$

as in KL-regularization (Todorov, 2009, 2006; Schulman et al., 2017; Galashov et al., 2019), but in the absence of any external rewards. In this case, one obtains an equation identical to (4.8) where the coefficients w_{ik} are simply replaced by $\pi_0(a_k|s_i)$, one to one. This apparently minor variation undercovers a major qualitative difference between absolute and relative action entropy objectives: as $\sum_k w_{ik} \geq 1$, absolute entropy-seeking favors visiting states with a large action accessibility, that is, where the sum $\sum_k w_{ik}$ and thus the argument of Eq. (4.8) tends to be largest. In contrast, as $\sum_k \pi_0(a_k|s_i) = 1$, maximizing relative entropies provides no preference for states s with large number of accessible actions $|\mathcal{A}(s)|$. This happens even if the default policy is uniform in the actions, as then the immediate intrinsic return becomes $-D_{\mathrm{KL}}(\pi(a|s)||\pi_0(a|s)) = \mathcal{H}(A|s) - \ln|\mathcal{A}(s)|$, instead of $\mathcal{H}(A|s)$. The negative logarithm penalizes visiting states with large number of actions, which is the opposite goal to occupying action-state path space.

4.3 Results

4.3.1 Entropy-seeking agents quickly fill physical space

In very simple environments with high symmetry and little constraints, like open space, maximizing path occupancy amounts to performing a random walk that chooses at every step any available action with equal probability. However, in realistic environments where space is not homogeneous or there are energetic limitations for moving, a random walk is no longer optimal. To illustrate how interesting behaviors arise from the POM principle in these cases, we first tested how an H agent moving in a 4-room and 4-food-sources environment (Fig. 4.2a) compares in occupying physical space to a random walker (RW) and to a reward seeking agent (R agent). The three agents are identical in most ways. They have nine possible movement actions, including not moving; they all have an internal state corresponding to the available energy, which reduces one unit at every time step and gets increased by a fixed amount (food gain) whenever a food source is visited; and they can move as long as their energy is non-zero. The total state space is the Cartesian product between physical space and internal energy. The agents differ however in their objective function. The H agent has a reward-free objective and implements POM by maximizing path action entropy, Eq. (4.2). In contrast, the R agent maximizes future discounted reward (in this case, food), and displays stochastic behavior through an ϵ-greedy action selection, with ϵ matched to the survival of the H agent (Supplemental Sec. 4.5.5 and Fig. 4.7a).

We find that the H agent generates behaviors that can be dubbed goal-directed and curiosity-driven (Video 1). First, by storing enough energy in its reservoir, the agent reaches far, entering the four rooms in the long term (Fig. 1b, left panel), and visiting every location of the arena except when food gain is small (Fig. 1c, blue line). In contrast, the R agent lingers over one of the food sources for most of the time (Fig. 1b, middle panel; Video 1). Although its ϵ-greedy action selection allows for brief exploration of other rooms, the R agent does not on average visit the whole arena (Fig. 1c, orange line). Finally, the random walker dies before it has time to visit a large fraction of the physical space (Fig. 1b, right panel). These

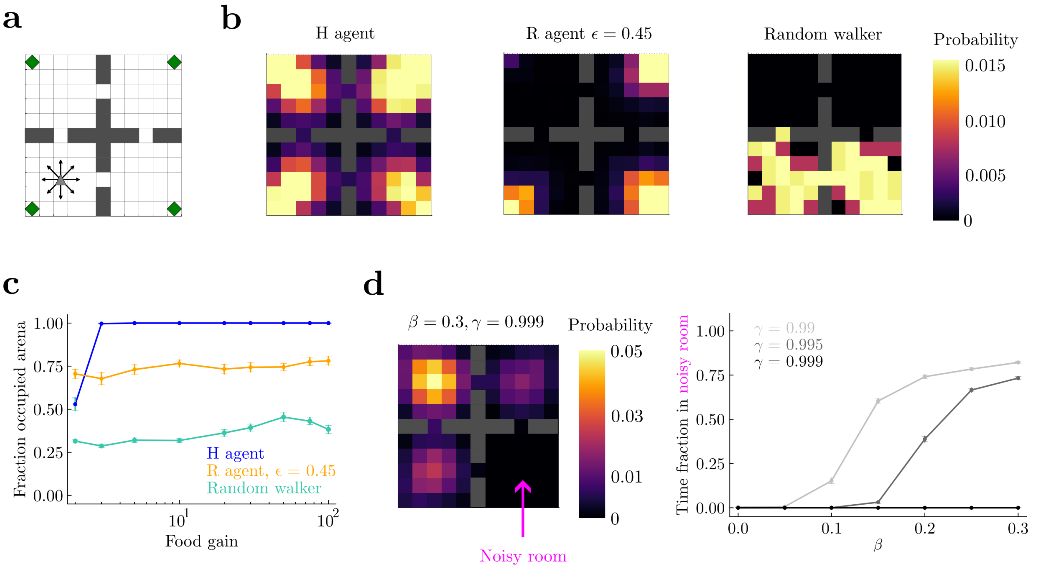


Figure 4.2: Maximizing future path occupancy leads to high occupancy of physical space. (a) Grid-world arena. The agents have nine available actions (arrows, and staying still) when alive (internal energy larger than zero) and away from walls. There are four rooms, each with a small food source in a corner (green diamonds). (b) Probability of visited spatial states for an entropy-seeking (H) agent, an ϵ-greedy reward (R) agent that survives as long as the H agent, and a random walker. Food gain = 10 units, maximum reservoir energy = 100, episodes of 5×10^4 time steps, and $(\alpha, \beta) = (1, 0)$ for the H agent. All agents are initialized in the middle of the lower left room. (c) Fraction of locations of the arena visited at least once per episode as a function of food gain. Error bars correspond to s.e.m over 50 episodes. (d) Noisy room problem. The bottom right room of the arena was noisy, such that agents in this room jump randomly to neighboring locations regardless of their actions. Food gain equals maximum reservoir energy = 100. Histogram of visited locations for an episode as long as in (b) for a H agent with $\beta = 0.3$ (left) and time fraction spent in the noisy room (right) show that H agents with $\beta > 0$ can either be attracted to the room or repelled depending on γ.

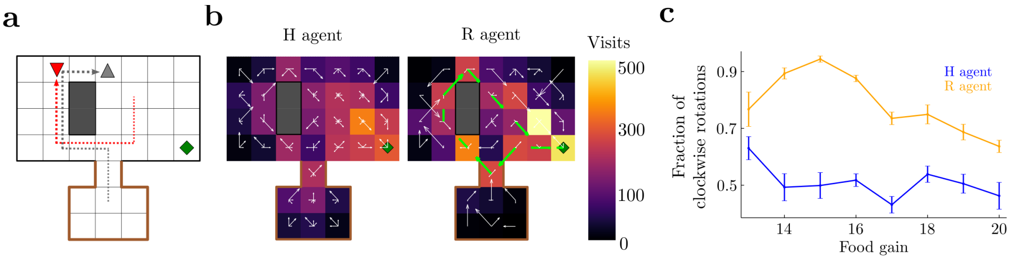


Figure 4.3: Complex hide-and-seek and escaping strategies in a prey-predator example. (a) Grid-world arena. The agent has nine available actions when alive and far from walls. There is a small food source in a corner (green diamond). A predator (red, down triangle) is attracted to the agent (gray, up triangle), such that when they are at the same location, the agent dies. The predator cannot enter the locations surrounded by the purple border. Arrows show a clockwise trajectory. (b) Histogram of visited spatial states across episodes for the H and R agents. The vector field at each location indicates probability of transition at each location. Green arrows on R agent show major motion directions associated with the its dominant clockwise rotation. (c) Fraction of clockwise rotations (as in panel (a)) to total rotations as a function of food gain, averaged over epochs of 500 timesteps. Error bars are s.e.m.

differences hold for a large range of food gains (Fig. 1c).

We next considered a slightly more complex environment where actions in one of the rooms lead to uniformly stochastic transitions to any of the neighboring locations. An H agent with $\beta > 0$ (see Eq. (4.2)) has a preference for stochastic state transitions, and *a priori* it could get attracted and stuck in the noisy room, where actions do not have any predictable effect –a spatial version of the noisy TV problem (Schmidhuber, 1991a; Burda et al., 2019). However, H agents also care about future states, and thus getting stuck in regions where energy cannot be predictably obtained will be avoided by sufficiently long-sighted agents (dependence on γ in Fig. 4.2d; Supplemental Sec. 4.5.5). This shows how H agents can tradeoff immediate with future action-state occupancy.

4.3.2 Hide and seek in a prey-predator interaction

More interesting behaviors arise from the POM principle in increasingly complex environments. To show this, we next considered a prey and a predator in a grid world with a safe area (house) and a single food source (Fig. 4.3a). The prey (a “mouse”, gray up triangle) is the agent whose behavior is optimized by maximizing future action path entropy, while the predator (a “cat”, red down triangle) acts passively chasing the prey. The state of the agent consists on its location and energy level, but it also includes the predator’s location being accurately perceived. The prey can move as in the previous 4-room grid world and has an energy reservoir as in the previous example. For simplicity, we only considered a food gain equal to the size of the energy reservoir, such that the agent fully replenishes its reservoir each time it visits the food source. The predator has the same available actions as the agent and is attracted to it stochastically, i.e. actions that move the predator towards the agent are more probable than those that move it away from it (Supplemental Sec. 4.5.5).

The entropy-seeking agent generates complex behaviors, not limited to visiting the food source to increase the energy buffer and hide at home. In particular, the agent very often first teases the cat and then performs a clockwise rotation around the obstacle, which forces the cat to chase it around, leaving the food source free for harvest (Fig. 4.3a, arrows show an example; Video 2, H agent). Importantly, this behavior is not restricted to clockwise rotations, as the agent performs an almost equal number of counterclockwise rotations to free the food area (Fig. 4.3c, H agent, blue line). The variability of these rotations in the entropy-seeking agent are manifest in the lack of virtually any preferred directionality of movement in the arena at any single position. Indeed, arrows pointing toward several directions indicate that on average the prey moves following different paths to get to the food source (Fig. 4.3b, H agent).

The behavior of the H agent was compared with an R agent that receives a reward of one each time it is alive and zero otherwise. To promote variable behavior in this agent as well, we implemented an ϵ-greedy action selection (Supplemental Sec. 4.5.5), where ϵ was chosen to match the expected lifetime of the H agent (Supplemental Fig. 4.7b). The behavior of the R agent was strikingly less variable than that of the H agent, spending more time close to the food source (Fig. 4.3b, R agent). Most importantly, while the H agent performs an almost equal number of clock and counterclockwise rotations, the R agent strongly prefers the clockwise rotations, reaching 90% of all observed rotations (Video 3, R-agent; Fig. 4.3c, orange line). This shows that the R agent mostly exploits only one strategy to survive and displays a smaller behavioral repertoire than the H agent.

4.3.3 Dancing in an entropy-seeking cartpole

In the previous examples, complex behaviors emerge as a consequence of the presence of obstacles, predators and limited food sources, but the actual dynamics of the agents are very coarse-grained. Here, we considered a system with physically realistic dynamics, the balancing cartpole (Barto et al., 1983; Florian, 2007), composed of a moving cart with an attached pole free to rotate (Fig. 4.4a). The cartpole is assumed to reach an absorbing state when either it hits a border, or when the pole angle exceeds 36 degrees. Thus, we consider a broad range of angles that makes the agents reach a larger state space than in standard settings (Brockman et al., 2016). We discretized the state space and used a linear interpolation to solve

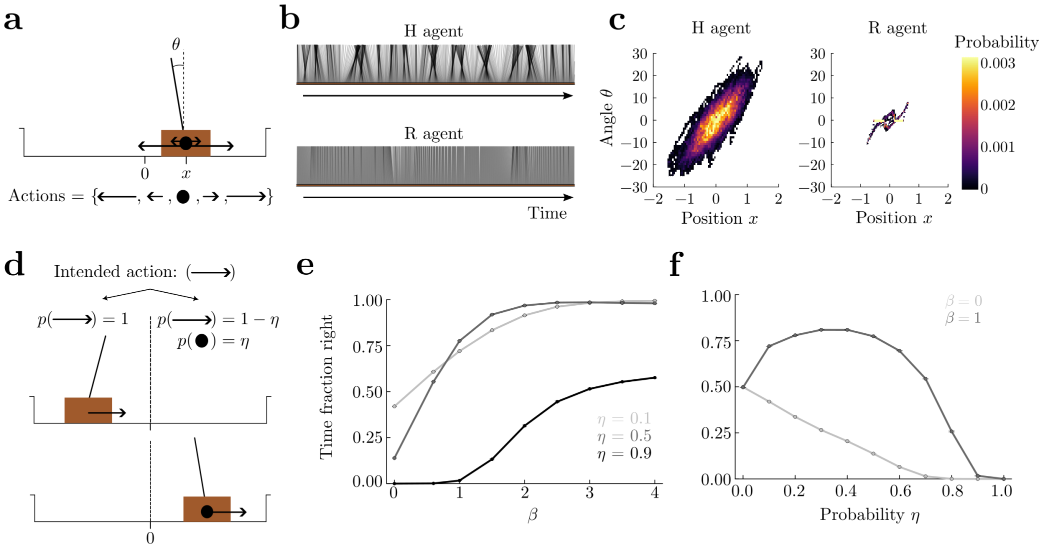


Figure 4.4: Dancing of an entropy-seeking cartpole. (a) The cart (brown rectangle) has a pole attached. The cartpole reaches an absorbing state if the magnitude of the angle θ exceeds $36\,\mathrm{deg}$ or its position reaches the borders. There are 5 available actions when alive: a big and a small force to either side (arrows on cartpole), and doing nothing (full circle). (b) Time-shifted snapshots of the pole in the reference frame of the cart as a function of time for the H (top) and R (bottom) agents. (c) Position and angle occupation for a 2×10^5 time step episode. (d) Here, the right half of the arena is stochastic, while the left remains deterministic. In the stochastic half, the intended state transition due to an applied action (force) succeeds with probability $1 - \eta$ (and thus zero force is applied with probability η). (e) Fraction of time spent on the right half of the arena increases as a function of β, regardless of the failure probability η. (f) The fraction has a non-monotonic behavior as a function of η when state entropy is important for the agent ($\beta = 1$), highlighting a stochastic resonance behavior. When the agents do not seek state entropy ($\beta = 0$) the fraction of time spent by the agent on the right decreases with the failure probability, and thus they avoid the stochastic right side. $\gamma = 0.99$ for panels (e,f).

for the optimal value function in Eq. (4.4), and to implement the optimal policy in Eq. (4.7), (Supplemental Sec. 4.5.5). The H agent widely occupies the horizontal position, and more strikingly it produces a wide variety of pole angles, constantly swinging sideways as if it were dancing (Video 4, H agent; Fig. 4.4b,c).

We compared the behavior of the H agent with that of an R agent that receives a reward of one for being alive and zero otherwise. As expected, the R agent maintains the pole close to the balanced position throughout most of a long episode (Fig. 4.4b, bottom), producing very little behavioral variability (Fig. 4.4c, right panel) and no movement that could be dubbed 'dancing' (Video 4, R agent). Although both agents use a similar strategy which keeps the pole pointing towards the center as much as possible (Fig. 4.4c, positive angles correlate with positive positions in both panels), the behavior of the R agent is qualitatively different, and is best described as a bang-bang sort of control for which the angle is kept very close to zero while the cart is allowed to travel and oscillate around the origin, which is more apparent in the actual paths of the agent (see trajectories in phase space in Video 5). We also find that the R agent does not display much variability in state space even after using an ϵ-greedy action selection (Supplemental Fig. 4.8), with ϵ chosen to match average lifetimes between agents (Supplemental Fig. 4.7c). This result showcases that the H agent exhibits the most appropriate sort of variability for a given average lifetime.

We finally introduced a slight variation to the environment, where the right half of the arena has stochastic state transitions. Here, when agents choose an action (force) to be executed, a state transition in the desired direction occurs with probability $1 - \eta$, and a transition corresponding to zero force occurs with probability η (Fig. 4.4d). Therefore, an H agent that seeks state entropy ($\beta > 0$) will show a preference for the right side, where there is in principle higher state entropy resulting from the stochastic transitions over more successor states than on the left side. Indeed, we find that H agents spend more time on the right side as β increases, regardless of the probability η (Fig. 4.4e). For fixed γ, spending more time on the right side can bring the life expectancy to decrease significantly depending on β and η (Supplemental Fig. 4.7 d-e). Interestingly, for $\beta > 0$ there is an optimal value of the noise η that maximizes the fraction of time spent on the right side (Fig. 4.4f), which is a form of stochastic resonance. Therefore, for different β, different qualitative behaviors emerge as a function of the noise level η.

4.3.4 Entropy-seeking agents can also seek entropy of others

Finally, we consider an example where an agent seeks to occupy path space, which includes another agent's location as well as its own. The agent can freely move (Fig. 4.5a; grey triangle) and open or close a fence by pressing a lever in a corner

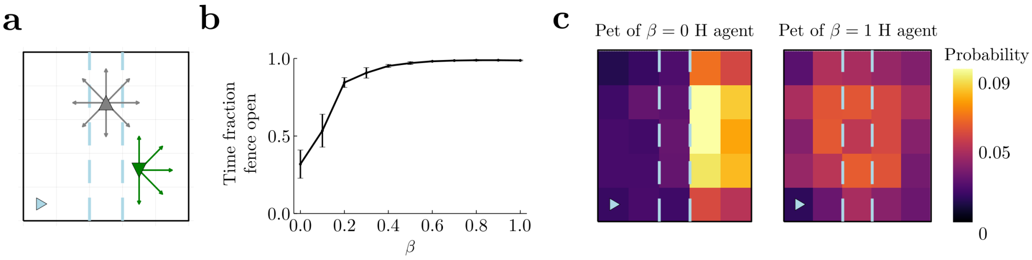


Figure 4.5: Modelling altruism through an optimal tradeoff between own action entropy and other's state entropy. (a) An agent (gray up triangle) has access to nine movement actions (gray arrows and doing nothing), and open or close a fence (dashed blue lines). This fence does not affect its movements. A pet (green, down triangle) has access to the same actions, and chooses one randomly at each timestep, but is constrained by the fence when closed. Pet location is part of the state of the agent. (b) As β in Eq. (4.2) is increased, the agent tends to leave the fence open for a larger fraction of time. This helps its pet reach other parts of the arena. Error bars correspond to s.e.m. (c) Occupation heatmaps for 2000 timestep-episodes for $\beta = 0$ (left) and $\beta = 1$ (right). In all cases $\alpha = 1$.

(blue triangle). The pet of the agent (green triangle) can freely move if the fence is open, but when the fence is closed the pet is confined to move in the region where it is currently located. The pet moves randomly at each step, but its available actions are restricted by its available space (Supplemental Sec. 4.5.5).

To maximize action-state path entropy, the agent ought to trade off the state entropy resulting from letting the pet free with the action entropy resulting from using the open-close action when visiting the lever location. The optimal tradeoff depends on the relative strength of action and state entropies. In fact, when state entropy weighs as much as action entropy ($\alpha = \beta = 1$), the fraction of time that the agent leaves the fence open is close to 1 (rightmost point in Fig. 4.5b) so that the pet is free to move (Fig. 4.5c, right panel; (A-S)-H agent). However, when the state entropy has zero weight ($\alpha = 1, \beta = 0$), the fraction of time that the fence remains open is close to 0.5 (leftmost point in Fig. 4.5b) and the pet remains confined on the right side for most of the time (Fig. 4.5c, left panel; (A)-H agent), the region where it was initially placed. As a function of β the fraction of time the fence is open increases. Therefore, the agent gives more freedom of its pet, as measured by the pet's state entropy, by curtailing its own action freedom, as measured by action entropy, thus becoming more "altruistic".

4.4 Discussion

Often, the success of agents in nature is not measured by the amount of reward obtained, but by their ability to expand in state space and perform complex behaviors. Here we have proposed that a major goal of intelligence is to ‘occupy path space’. External rewards are thus the means to move and occupy action-state path space, not the goal of behavior. In a MDP setting, we have shown that the intuitive notion of path occupancy is captured by future action-state path entropy, and we have proposed that behavior is driven by the maximization of this sole intrinsic goal –the POM principle. We have solved the associated Bellman equation and provided a convergent iterative map to determine the optimal policy.

In four examples we have shown that the POM principle, along with the agent’s constraints and dynamics, leads to complex behaviors that are not observed in other simple reward maximizing agents. Quick filling of physical space by a moving agent, hide-and-seek behavior and variable escaping routes in a predator-prey example, dancing in a realistic cartpole dynamical system and altruistic behavior in an agent-and-pet duo are all behaviors that strike as being playful, curiosity-driven and energetic. To the human eye, these behaviors look genuinely goal-directed (see SI Sec. 4.5.6). Although the agent does not have any externally designed goal, it still seeks the regions of state space (e.g. with resources) that allow it to move the longest in a way that maximizes future path action-state entropy.

A related set of algorithms, known as empowerment, have also proposed using reward-free objectives as the sole goal of behavior (Klyubin et al., 2005; Jung et al., 2011; Mohamed and Jimenez Rezende, 2015). In this approach, the mutual information between a sequence of actions and the final state is maximized. This makes empowerment agents prefer states where actions leads to large and predictable changes, such as unstable fixed points (Jung et al., 2011). One drawback is that empowered agents tend to remain close to those states without producing diverse behavioral repertoires, as it also happens in causal entropy approaches (Wissner-Gross and Freer, 2013). For instance, in the cartpole setting, both empowered and causal entropy agents balance the pole upwards and cease variable behavior when that state is reached (Jung et al., 2011; Wissner-Gross and Freer, 2013). Another difference is that empowerment is not additive over paths, and thus it cannot be formalized as a cumulative per-step objective (Supplemental Sec. 4.5.7) (Jung et al., 2011; Leibfried et al., 2019; Mohamed and Jimenez Rezende, 2015; Volpi and Polani, 2020), in contrast to action-state path entropy. We note, however, that an approximation to empowerment having the desired additive property could be obtained from our framework by putting $\beta < 0$ in Eq. (4.2), such that more predictable state transitions are preferred. Other reward-free RL settings and pure exploration objectives have been proposed in the past (Hazan et al., 2019; Lee et al., 2019; Jin et al., 2020; Zhang et al., 2021a; Mutti and Restelli, 2020; Mutti et al., 2021; Pathak et al., 2017; Eysenbach and Levine, 2021), but this body of work typically investigates how to efficiently sample MDPs to construct near-optimal policies when reward functions are introduced in the exploitation phase. More importantly, this work differs from ours in that the POM principle generates goal-directedness and behavioral variability even in known environments (see examples above).

Several steps remain to have a more complete POM theory. One is to study learning in environments where state transitions are not known. Previous related attempts have introduced Z-learning (Todorov, 2006, 2009) and G-learning (Fox et al., 2015) using off-policy methods, so our results could be extended to learning following similar lines. Other possibilities are using transition estimators using counts or pseudo-counts (Bellemare et al., 2016), or hashing (Tang et al., 2017), for the learning of the transition matrices. One potential advantage of our framework is that, as entropy-seeking behavior obviates external rewards, those rewards do not need to be learned and optimized, and thus the learning problem reduces to transition matrices learning. In addition, modeling and injecting prior information could be particularly simple in our setting in view that intrinsic entropy rewards can be easily bounded before the learning process if action space is known. Therefore, initializing the state-value function to the lower or upper bounds of the action-state path entropy could naturally model pessimism or optimism during learning, respectively.

All in all, we have introduced POM as a novel theory of behavior, which promises new ways of understanding goal-directedness without reward maximization, and that can be applied to artificial agents to discover by themselves ways of surviving and occupying action-state space.

4.5 Appendix

4.5.1 Entropy measures the occupancy of action-state paths

We consider a time-homogeneous Markov decision process with finite state set $\mathcal{S}$ and finite action set $\mathcal{A}(s)$ for every state $s \in \mathcal{S}$. Henceforth, the action-state $x_j = (a_j, s_j)$ is any joint pair of one available action a_j and one possible successor state s_j that results from making that action under policy $\pi \equiv \{\pi(a|s)\}$ from the action-state $x_i = (a_i, s_i)$. By assumption, the availability of action a_j depends on the previous state s_i alone, not on a_i. Thus, the transition probability from x_i to x_j in one time step is $p_{ij} = \pi(a_j|s_i)p(s_j|s_i, a_j)$, where $p(s_j|s_i, a_i)$ is the conditional probability of transitioning from state s_i to s_j given that action a_j is performed. Although there is no dependence of the previous action a_i on this transition probability, it is notationally convenient to define transitions between action-states. We conceive of rational agents as maximizing future action-state path occupancy. Any measure of occupancy should obey the intuitive Conditions *1-4* listed below.

Intuitive Conditions for a measure of action-state occupancy:

1. *Occupancy gain of action-state x_j from x_i is a function of the transition probability p_{ij}, $C(p_{ij})$*

2. *Performing a low probability transition leads to a higher occupancy gain than performing a high probability transition, that is, $C(p_{ij})$ decreases with p_{ij}*

3. *The first order derivative $C'(p_{ij})$ is continuous for $p_{ij} \in (0, 1)$*

4. *(Definition: the action-state occupancy of a one-step path from action-state x_i is the expectation over occupancy gains of the immediate successor action-states, $C_i^{(1)} \equiv \sum_j p_{ij} C(p_{ij})$)*

   The action-state occupancy of a two-steps path is additive,

   $C_i^{(2)} \equiv \sum_{jk} p_{ij} p_{jk} C(p_{ij} p_{jk}) = C_i^{(1)} + \sum_j p_{ij} C_j^{(1)}$

   for any choice of the p_{ij} and initial x_i

Condition *1* simply states that occupancy gain from an initial action-state is defined over the transition probabilities to successor action-states in a sample space. Condition *2* implies that performing a low probability transition leads to a higher occupancy of the successor states than performing a high probability transition. This is because performing a rare transition allows the agent to occupy a space that was left initially unoccupied. Condition *3* imposes smoothness of the measure.

In Condition *4* we have defined the occupancy of the successor action-states (one-step paths) in the Markov chain as the expected occupancy gain. Condition *4* is the central property, and it imposes that the occupancy of action-states paths with two steps can be broken down into a sum of the occupancies of action-states at each time step. Note that the action-state path occupancy can be written as

$$\begin{aligned} C_i^{(2)} \equiv \sum_{jk} p_{ij} p_{jk} C(p_{ij} p_{jk}) &= \sum_{j} p_{ij} C(p_{ij}) + \sum_{jk} p_{ij} p_{jk} C(p_{jk}) \\ &= \sum_{jk} p_{ij} p_{jk} \left(C(p_{ij}) + C(p_{jk}) \right), \end{aligned}$$

which imposes a strong condition on the function $C(p)$. Note also that the sum $\sum_{jk} p_{ij} p_{jk} C(p_{ij} p_{jk})$ extends the notion of action-state to a path of two consecutive action-states, each path having probability $p_{ij} p_{jk}$ due to the (time-homogeneous) Markov property. The last equality is an identity. While here we consider paths of length equal to 2, further below we show that there is no difference in imposing additivity to paths of any fixed or random length (Corollary 2).

Theorem 1. *$C(p) = -k \ln p$ with $k > 0$ is the only function that satisfies Conditions* 1-4

Corollary 1. *The entropy $C_i^{(1)} = -k \sum_j p_{ij} \ln p_{ij}$ is the only measure of action-state occupancy of successor action-states x_j from x_i with transition probabilities p_{ij} consistent with Conditions* 1-4.

Proof. Put $p_{1,1} = 1$ and $p_{1,j} = 0$ for $j \neq 1$. Then, Condition *4* reads $C(1) = C(1) + C(1)$ when the initial action-state is x_1, which implies $C(1) = 0$.

Now, take a Markov chain with $p_{0,0} = 1$, $p_{1,0} = 1 - t > 0$, $p_{1,2} = t > 0$, $p_{2,0} = p_{2,1} = 0$, $p_{2,j} = 1/n$ for $j = 3, ..., n+2$ and $n > 0$, and $p_{k,0} = 1$ for $k = 3, ..., n+2$. In this chain, the state 0 is absorbing and all others are transient (here action-states are simply referred to as states). Starting from state 1, transition to the transient state 2 happens with probability t and to the absorbing state 0 with probability $1 - t$. From state 2 a transition to states $j = 3, ..., n+2$ happens with equal probability. From any of those states, a deterministic transition to 0 ensues. (These last transitions can only happen in the third time step, and although it will be relevant later on, it is no used in the current proof, which focuses on paths of length two.) Then, Condition *4* with initial state 1 reads $tC(t/n) + (1-t)C(1-t) = tC(t) + (1-t)C(1-t) + tC(1/n) + (1-t)C(1)$, and hence $C(t/n) = C(t) + C(1/n)$ for any $0 < t < 1$ and integer $n > 0$. By Condition *3* and taking derivative with respect to t in both sides, we obtain $C'(t/n) = nC'(t)$, and multiplying in both sides by t we obtain $\frac{t}{n}C'(\frac{t}{n}) = tC'(t)$. By replacing t with nt, we get $tC'(t) = ntC'(nt)$, provided that $nt < 1$.

We will now show that $tC'(t)$ is constant. In the last equation replace t by t/m by integer $m > 0$ to get the last equivalence in $tC'(t) = \frac{t}{m}C'(\frac{t}{m}) = \frac{n}{m}tC'(\frac{n}{m}t)$ (the first equivalence is obvious). These equivalences are valid for positive $t < 1$ and $\frac{n}{m}t < 1$. Let $0 < s < 1$ and $n = \lfloor ms/t \rfloor$ be the largest integer smaller than ms/t. Therefore, as m increases $\frac{n}{m}t < 1$ and approaches s as close as desired. By Condition *3* the function $xC'(x)$ is continuous, and therefore $\lim_{m\to\infty} \frac{n}{m}tC'(\frac{n}{m}t) = sC'(s)$. The basic idea is that we can first compress t as much as needed by the integer factor m and then expand it by the integer factor n so that nt/m is as close as desired to s. This shows that $sC'(s) = tC'(t)$ for $s, t \in (0, 1)$, and therefore $tC'(t)$ is constant.

Assume that $tC'(t) = -k$. Then, by integrating we obtain $C(t) = -k \ln t + a$, but $a = 0$ due to $C(1) = 0$, and $k > 0$ due to Condition *2*. Together with the above, we can now proof the theorem by noticing that the solution satisfies Condition *4* for any choice of the p_{ij}. □

Corollary 2. Condition *4* can be replaced by an equivalent condition that requires additivity of paths of any finite length n with no change in the above proof. We first introduce some notation: the probability of path $i_0, i_1, ..., i_n$ is $p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n}$, where i_t refers to the state visited at step t and i_0 is the initial state. Then the new Condition *4* reads in terms of the action-state occupancy of paths of length n as

$$\begin{aligned} C_{i_0}^{(n)} &= \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} C\left(p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n}\right) \\ &= \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_1,i_2}) + ... \\ &+ \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} C\left(p_{i_{n-1},i_n}\right) \\ &= \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} \left(C(p_{i_0,i_1}) + C(p_{i_1,i_2}) ... + C(p_{i_{n-1},i_n})\right) \ , \end{aligned}$$

for any time-homogeneous Markov chain. By choosing the particular chains used in Theorem 1, we arrive again to the same unique solution $C(p) = -k \ln p$ after using $C(1) = 0$ repeated times, which obviously solves the above equation for any chain and length path. Indeed, note that for the second chain in Theorem 1, from initial state 1 the absorbing state is reached in three time steps with probability one, and thus the above sum contains all $C(1)$ starting from the third terms, which contribute zero to the sum.

The above entropy measure of action-state path occupancy can be extended to the case where there is a discount factor $0 < \gamma < 1$. To do so, we assume now that the paths can have a random length $n \geq 1$ that follows a geometric distribution, $p_n = \gamma^{n-1}(1-\gamma)$. In this case, the occupancy of the paths is

$$\begin{aligned} C_{\text{global}} &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) \\ &\quad +\gamma^2(1-\gamma)\sum_{i_1,i_2,i_3} p_{i_0,i_1}p_{i_1,i_2}p_{i_2,i_3} C(p_{i_0,i_1}p_{i_1,i_2}p_{i_2,i_3}) + \ldots \end{aligned} \tag{4.9}$$

where the n-th term in the sum is the expected occupancy gain of paths of length n weighted by the probability of a having a path with exactly such a length.

Equivalently, a path in course can grow one step further with probability γ or be extinguished with probability $1-\gamma$. Therefore, the occupancy in Eq. (4.9) should also be equal to the sum of the expected occupancy gains of the local states along the paths, defined as

$$C_{\text{local}} = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_1,i_2}) + \gamma^2 \sum_{i_1,i_2,i_3} p_{i_0,i_1}p_{i_1,i_2}p_{i_2,i_3} C(p_{i_2,i_3}) + \ldots \tag{4.10}$$

where the first term is the expected occupancy gain given by the initial condition, the second term is the expected occupancy gain in the next step weighted by the probability of having a path length of at least two steps, and so on.

Eqs. (4.9-4.10), after using the Markov chain in Corollary 2, reduce to

$$\begin{aligned} C_{\text{global}} &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) \\ &\quad +\gamma^2(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) + \ldots \\ &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) \end{aligned}$$

and

$$C_{\text{local}} = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_1,i_2}),$$

where we have used $p_{i_2,i_3} = 1$ because all transitions in the third step are deterministic.

Equality of these two quantities leads to Condition *4*, specifically,

$$\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_1,i_2}).$$

Therefore, the only consistent measure of occupancy with temporal discount is the entropy. Obviously, the equality of global and local time-discounted occupancies measured by entropy holds for any time-homogeneous or inhomogeneous Markov chain.

4.5.2 Critical policies and critical state-value functions

Here, the expected return following policy π in Eq. (4.10), known as the state-value function, is written recursively using the Bellman equation. Then, we find a non-linear system of equations for the critical policy and critical state-value function by taking partial derivatives with respect to the policy probabilities (Theorem 2).

Using Eq. (4.10) and Theorem 1 with $k = 1$, we define the expected return from state s under policy π as

$$V_\pi(s) = -\sum_{i_1} p_{s,i_1} \ln p_{s,i_1} - \gamma \sum_{i_1,i_2} p_{s,i_1} p_{i_1,i_2} \ln p_{i_1,i_2} - \gamma^2 \sum_{i_1,i_2,i_3} p_{s,i_1} p_{i_1,i_2} p_{i_2,i_3} \ln p_{i_2,i_3} + \ldots \tag{4.11}$$

where p_{s,i_1} is the transition probability from state s to action-state $x_{i_1} = (a_{i_1}, s_{i_1})$. Note that in Eq. (4.10) we have replaced the initial action-state i_0 by the initial state s alone, as the previous action that led to it does no affect the transition probabilities in the Markov decision process setting. The expected returns satisfy the standard recurrence relationship (Sutton et al., 1998)

$$\begin{aligned} V_\pi(s) &= \sum_{a,s'} p_{s,(a,s')} \left(-\ln p_{s,(a,s')} + \gamma V_\pi(s')\right) \\ &= \sum_{a,s'} \pi(a|s) p(s'|s,a) \left(-\ln \pi(a|s) p(s'|s,a) + \gamma V_\pi(s')\right). \end{aligned} \tag{4.12}$$

Here, we have unpacked the sum over the action-state i_1 into a sum over (a, s'), where a is the action made in state s and s' is its successor. The second equation shows, in a more standard notation, the explicit dependence of the expected return on the policy. It also highlights that the intrinsic immediate reward takes the form $R_{\text{intrinsic}}(s, a, s') = -\ln \pi(a|s)p(s'|s, a)$, which is unbounded.

From Eq. (4.11) it is easy to see that the expected return exists (is finite) for any policy π if the Markov decision process has a finite number of actions and states. Due to the properties of entropy, Eq. (4.11) is a sum of non-negative numbers bounded by $H_{max} = \ln(|A|_{max}|S|)$ ($|A|_{max}$ is the maximum number of available actions from any state) weighted by the geometric series, which guarantees convergence of the infinite sum for $-1 < \gamma < 1$. An obvious, but relevant, implication of the above is that the expected return is non-negative and bounded, $0 \leq V_\pi(s) \leq H_{max}/(1-\gamma)$, for any state and policy.

While in Eq. (4.12) the immediate intrinsic reward is the sum of the action and state occupancies, $R_{\text{intrinsic}}(s,a,s') = -\ln \pi(a|s)p(s'|s,a) = -\ln \pi(a|s) - \ln p(s'|s,a)$, we can generalize this reward to consider any weighted mixture of entropies as $R_{\text{intrinsic}}(s,a,s') = -\alpha \ln \pi(a|s) - \beta \ln p(s'|s,a)$ for any two numbers $\alpha > 0$ and $\beta \geq 0$. In particular, for $(\alpha,\beta) = (1,1)$ we recover the action-state occupancy of Eq. (4.12), and for $(\alpha,\beta) = (1,0)$ and $(\alpha,\beta) = (0,1)$ we only consider action or state occupancy, respectively. The case $(\alpha,\beta) = (0,1)$ is understood as the limit case where α becomes infinitely small. We note that the case $(\alpha,\beta) = (1,0)$ has often been used along with an external reward with the aim of regularizing the external reward objective (Ziebart, 2010; Todorov, 2009; Schulman et al., 2017; Haarnoja et al., 2018; Hausman et al., 2018). We also note that the case $(\alpha,\beta) = (1,-1)$, with negative β, constitutes an approximation to empowerment (Klyubin et al., 2005; Jung et al., 2011): the agent tries to maximize action entropy while minimizing state entropy conditioned to the previous action-state, which favors paths where there is more control on the resulting states. However, we do not consider this case in this paper.

Under the more general intrinsic reward, the expected return obeys

$$V_\pi(s) = \sum_{a,s'} \pi(a|s)p(s'|s,a)\left(-\ln \pi^\alpha(a|s)p^\beta(s'|s,a) + \gamma V_\pi(s')\right). \tag{4.13}$$

Our goal is to maximize the expected return over the policy probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$ to obtain the optimal policy. Note that for $\alpha > 0$ and $\beta \geq 0$ the expected return is non-negative, $V_\pi(s) \geq 0$.

Theorem 2. *The critical values $V^c(s)$ of the expected returns $V_\pi(s)$ in equation (4.13) with respect to the policy probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$ obey*

$$V^c(s) = \alpha \ln Z(s) = \alpha \ln \left[\sum_{a \in \mathcal{A}(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a)V^c(s')\right)\right] \tag{4.14}$$

where $\mathcal{H}(S'|s,a) = -\sum_{s'} p(s'|s,a)\ln p(s'|s,a)$ is the entropy of the successors of s after performing action a, and $Z(s)$ is the partition function.

The critical points (critical policies) are

$$\pi^c(a|s) = \frac{1}{Z(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a)V^c(s')\right), \tag{4.15}$$

one per critical value, where the partition function $Z(s)$ is the normalization constant.

Defining $z_i = \exp(\alpha^{-1}\gamma V^c(s_i))$, $p_{ijk} = p(s_j|s_i, a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i, a_k)$, Eq. (4.14) can be compactly rewritten as

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \tag{4.16}$$

where the matrix with coefficients $w_{ik} \in \{0, 1\}$ indicates whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$), and j extends over all states, with the understanding that if a state s_j is not a possible successor from state s_i and action a_k then $p_{ijk} = 0$.

Note that the we simultaneously optimize $|S|$ expected returns, one per state s, each with respect to the set of probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$.

Proof. We first note that the expected return in Eq. (4.2) is continuous and has continuous derivatives with respect to the policy except at the boundaries (i.e., $\pi(a|s) = 0$ for some action-state (a, s)). Choosing a state s, we first take partial derivatives with respect to $\pi(a|s)$ for each $a \in \mathcal{A}(s)$ in both sides of (4.13), and then evaluate them at a critical point π^c to obtain the condition

$$\begin{aligned} \lambda(s, s) &= \sum_{s'} p(s'|s, a) \left(-\ln(\pi^c(a|s))^\alpha p^\beta(s'|s, a) + \gamma V^c(s')\right) - \alpha \\ &\quad + \gamma \sum_{b,s'} \pi^c(b|s) p(s'|s, b) \lambda(s', s) \qquad (4.17) \\ &= -\alpha \ln \pi^c(a|s) - \beta \sum_{s'} p(s'|s, a) \ln p(s'|s, a) - \alpha \\ &\quad + \gamma \sum_{s'} p(s'|s, a) V^c(s') + \gamma \sum_{b,s'} \pi^c(b|s) p(s'|s, b) \lambda(s', s), \qquad (4.18) \end{aligned}$$

where we have defined the partial derivative at the critical point $\frac{\partial V_\pi(s')}{\partial \pi(a|s)}|_{\pi^c} \equiv \lambda(s', s)$ and used the fact that this partial derivative should be action-independent. To understand this, note that the critical policy should lie in the simplex $\sum_a \pi(a|s) = 1$, $\pi(a|s) \geq 0$, and therefore the gradient of $V_\pi(s')$ with respect to the $\pi(a|s)$ at the critical policy should be along the normal to the constraint surface, i.e., the diagonal direction (hence, action-independent), or be zero. Indeed, the action-independence of the $\lambda(s', s)$ also results from interpreting them as Lagrange multipliers: $\lambda(s', s)$ is the Lagrange multiplier corresponding to the state-value function at s', $V_\pi(s')$, associated to the constraint $\sum_a \pi(a|s) = 1$, $\pi(a|s) \geq 0$, defining the simplex where the probabilities $\{\pi(a|s) : a \in A(s)\}$ lie.

Noticing that the last term of Eq. (4.18) does not depend on a, we can solve for the critical policy $\pi^c(a|s)$ to obtain equation (4.15). Eq. (4.15) implicitly relates the critical policy with the critical value of the expected returns from each state s. Inserting the critical policy (4.15) into Eq. (4.13), we get (4.14), which is an implicit non-linear system of equations exclusively depending on the critical values.

It is easy to verify that the partial derivatives of $V_\pi(s)$ in Eq. (4.13) with respect to $\pi(a'|s')$ for $s \neq s'$ are

$$\lambda(s,s') = \gamma \sum_{s''} p(s''|s)\lambda(s'',s'),$$

and thus they provide no additional constraint on the critical policy. [1]

□

We finally show that the optimal expected returns, as defined from the Bellman optimality equation

$$V^*(s) = \max_{\pi(\cdot|s)} \sum_{a,s'} \pi(a|s)p(s'|s,a)\left(-\ln \pi^\alpha(a|s)p^\beta(s'|s,a) + \gamma V^*(s')\right), \quad (4.19)$$

obey the same Eq. (4.14) as the critical values of Eq. (4.13) do. To see this, note that after taking partial derivatives with respect to $\pi(a|s)$ for each $a \in \mathcal{A}(s)$ on the right-hand side of Eq. (4.19) we get

$$0 = -\alpha \ln \pi(a|s) - \beta \sum_{s'} p(s'|s,a)\ln p(s'|s,a) + \gamma \sum_{s'} p(s'|s,a)V^*(s') - \alpha + \lambda(s), \quad (4.20)$$

where $\lambda(s)$ is the Lagrange multiplier associated to the constraint $\sum_a \pi(a|s) = 1$. This equation, except for the irrelevant action-independent Lagrange multipliers, is identical to Eq. (4.18). Eq. (4.14) follows from inserting the resulting optimal policy into the Bellman optimality equation.

[1] This set of equations along with Eq. (4.18) generates a linear system of $\mathcal{S}^2$ equations for the $\mathcal{S}^2$ unknowns $\lambda(s,s')$. In the next subsection we show that the critical values $V^c(s)$ and critical policy $\pi^c(a|s)$ exists and are unique, and thus the system of equations for $\lambda(s,s')$ is of the type $\Lambda = \gamma P^\intercal \Lambda + F$, with unique matrices $\Lambda_{ss'} = \lambda(s,s')$, $P_{s's} = p(s'|s) \equiv \sum_a \pi^c(a|s)p(s'|s,a)$ and $F_{s's}$ is a diagonal matrix with $F_{ss} = V^c(s) - \alpha$. Because P is a stochastic matrix, it does not have eigenvalues larger than one. Therefore the matrix $\mathbb{I} - \gamma P^\intercal$ with $\gamma < 1$ does not have zero eigenvalues, and thus it is invertible. The solution to the system is then unique and given thenby $\Lambda = (\mathbb{I} - \gamma P^\intercal)^{-1}F$.

4.5.3 Unicity of the optimal value and policy, and convergence of the algorithm

We now prove that the critical value $V^c(s)$ is unique, in other words, equation (4.14) admits a single solution (Theorem 3). We later prove that the solution is the optimal expected return (Theorem 4).

Theorem 3. *With the definitions in Theorem 2, the system of equations*

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \tag{4.21}$$

with $0 < \gamma < 1$, $\alpha > 0$ and $\beta \geq 0$ has a unique solution in the positive first orthant $z_i > 0$, provided that for all i there exists at least one k such that $w_{ik} = 1$. The solution satisfies $z_i \geq 1$.

Moreover, given any initial condition $z_i^{(0)} > 0$ for all i, the infinite series $z_i^{(n)}$ defined through the iterative map

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} \tag{4.22}$$

for $n \geq 0$ converges to a finite limit $z_i^{\infty} \geq 1$, and this limit is the unique solution of equation (4.21)

Note that the condition that for all i there exists at least one k such that $w_{ik} = 1$ imposes virtually no restriction, as it only asks for the presence of at least one available action in each state. For instance, in absorbing states, the action leads to the same state.

Importantly, proving that the map (4.22) has a single limit regardless of the initial condition in the positive first orthant $z_i^{(0)} > 0$ suffices to prove that equation (4.21) has a unique solution in that region, as then no other fix point of the map can exist. Additionally, since the solution is unique and satisfies $z_i^{\infty} \geq 1$, the critical state-value function that solves equation (4.14) is unique, and $V^c(s_i) = \alpha\gamma^{-1} \ln z_i^{\infty} \geq 0$, consistent with its properties.

The map (4.22) provides a useful value-iteration algorithm used in examples shown in the Results section, and empirically is found to rapidly converge to the solution.

Proof. We call the series $z_i^{(n)}$ with initial condition $z_i^{(0)} = 1$ for all i the *main* series. We first show that the main series is monotonic non-decreasing.

For $n = 1$, we get

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^\gamma \geq 1 = z_i^{(0)} \tag{4.23}$$

for all i, using that there exists k for which, $w_{ik} = 1$, w_{ik} is non-negative for all i and k, $\mathcal{H}_{ik} \geq 0$ and the power function x^γ is increasing with its argument.

Assume that for some $n > 0$, $z_i^{(n)} \geq z_i^{(n-1)}$ for all i. Then

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^\gamma \geq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n-1)} \right)^{p_{ijk}} \right)^\gamma = z_i^{(n)} \tag{4.24}$$

using the same properties as before, which proves the assertion for all n by induction.

Now let us show that the main series is bounded. Define $\mathcal{H}_{\text{max}} = \max_{ik} \mathcal{H}_{ik}$, and obviously $\mathcal{H}_{\text{max}} \geq 0$.

For $n = 1$ we have

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^\gamma \leq \left(|A|_{\text{max}} e^{\mathcal{H}_{\text{max}}} \right)^\gamma \equiv c^\gamma \tag{4.25}$$

(remember that $|A|_{\text{max}}$ is the maximum number of available actions from any state).

For $n = 2$,

$$\begin{aligned} z_i^{(2)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(1)} \right)^{p_{ijk}} \right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j c^{\gamma p_{ijk}} \right)^\gamma \\ &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} c^\gamma \right)^\gamma = c^{\gamma^2} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^\gamma \leq c^{\gamma+\gamma^2} \end{aligned}$$

using the standard properties, $\sum_j p_{ijk} = 1$ and Eq. (4.25).

Assume that for some $n > 1$ we have $z_i^{(n)} \leq c^{\gamma+\gamma^2+\ldots+\gamma^n}$. We have just showed that this is true for $n = 2$. Then

$$\begin{aligned} z_i^{(n+1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} c^{\gamma+\ldots+\gamma^n} \right)^\gamma \\ &= c^{\gamma^2+\ldots+\gamma^{n+1}} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^\gamma \leq c^{\gamma+\ldots+\gamma^{n+1}} \end{aligned}$$

and therefore it is true for all $n \geq 0$ by induction.

Therefore the series $z_i^{(n)}$ is bounded by $c^{1/(1-\gamma)}$. Together with the monotonicity of the series, we have now proved that the limit z_i^{∞} of the series exists. Moreover, $z_i^{\infty} \geq z_i^0 = 1$.

The above results can be intuitively understood: the 'all ones' initial condition of the main series corresponds to an initial guess of the state-value function equal to zero everywhere. The iterative map corresponds to state-value iteration to a more optimistic value: as intrinsic reward based on entropy is always non-negative, the z-values monotonically increase after every iteration. Finally, the z-values reach a limit because the state-value function is bounded.

We now show the central result that the series obtained by using the iterative map starting from any initial condition in the positive first orthant can be bounded below and above by two series that converge to the main series. Therefore, by building 'sandwich' series we will confirm that any other series has the same limit as the main series.

Let the $y_i^{(0)} = u_i > 0$ be the initial condition of the series $y_i^{(n)}$ obeying the iterative map (4.22), and define $u_{\text{min}} = \min_i u_i$ and $u_{\text{max}} = \max_i u_i$. Obviously, $u_{\text{min}} > 0$ and $u_{\text{max}} > 0$. Applying the iterative map once, we get

$$\begin{aligned} y_i^{(1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(0)} \right)^{p_{ijk}} \right)^{\gamma} \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(u_{\text{max}} \right)^{p_{ijk}} \right)^{\gamma} \\ &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} u_{\text{max}} \right)^{\gamma} = u_{\text{max}}^{\gamma} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^{\gamma} = u_{\text{max}}^{\gamma} z_i^{(1)} \end{aligned}$$

where in the last step we have used the values of the main series in the first iteration. We can similarly lower-bound $y_i^{(1)}$ to finally show that it is both lower- and upper-bounded by $z_i^{(1)}$ with different multiplicative constants,

$$u_{\text{min}}^{\gamma} z_i^{(1)} \leq y_i^{(1)} \leq u_{\text{max}}^{\gamma} z_i^{(1)} \tag{4.26}$$

Now, assume that

$$u_{\text{min}}^{\gamma^n} z_i^{(n)} \leq y_i^{(n)} \leq u_{\text{max}}^{\gamma^n} z_i^{(n)} \tag{4.27}$$

is true for some $n > 0$. Then, for $n + 1$ we get

$$\begin{aligned} y_i^{(n+1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(n)}\right)^{p_{ijk}}\right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(u_{\text{max}}^{\gamma^n} z_i^{(n)}\right)^{p_{ijk}}\right)^\gamma \\ &= u_{\text{max}}^{\gamma^{n+1}} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_i^{(n)}\right)^{p_{ijk}}\right)^\gamma = u_{\text{max}}^{\gamma^{n+1}} z_i^{(n+1)} \end{aligned}$$

by simply extracting the common factor in the fourth expression, remembering that $\sum_j p_{ijk} = 1$, and using the definition of the main series in the last one. By repeating the same with the lower bound, we finally find that (4.27) holds also for $n + 1$, and then, by induction, for every $n > 0$.

The proof concludes by noticing that the limit of both $u_{\text{max}}^{\gamma^n}$ and $u_{\text{min}}^{\gamma^n}$ is 1, and therefore using (4.27) the limit y_i^∞ of the series $y_i^{(n)}$ equals the limit of the main series, $y_i^\infty = z_i^\infty$.

Note that the iterative map (4.22) is not necessarily contractive in the Euclidean metric, as it is possible that, depending on the values of u_{min} and u_{max} and the changes in the main series, the bounds in Eq. (4.27) initially diverge to finally converge in the limit.

□

Theorem 4. *The (unique) critical value $V^c(s)$ is the optimal expected return, that is, the one that attains the maximum expected return at every state for any policy, and we write $V^c(s) = V^*(s)$*

Proof. To show that $V^c(s)$ is the optimal expected return, we note that the maximum of the functions $V_\pi(s)$ with respect to policy π should be at the critical policy or at the boundaries of the simplices defined by $\sum_a \pi(a|s) = 1$ with $0 \leq \pi(a|s) \leq 1$ for every a and s, as the expected return $V_\pi(s)$ is continuous and has continuous derivatives with respect to the policy except at the boundaries. At the policy boundary, there exists a non-empty subset of states s_i and a non-empty set of actions a_k for which $\pi(a_k|s_i) = 0$. Computing the critical value of the expected return along that policy boundary is identical to moving from the original to a new problem where we replace the graph connectivity matrix w_{ik} in Eq. (4.21) by a new one v_{ik} such that $v_{ik} \leq w_{ik}$ (remember that at the boundary there should be an action a_k that were initially available from state s_i, $w_{ik} = 1$, that at the policy boundary is forbidden, $v_{ik} = 0$). We now define the convergent series $z_i^{(n)}$ and $y_i^{(n)}$ for the original and new problems respectively by using the iterative map (4.22) with initial conditions equal to all ones. We prove now that $z_i^{(n)} \geq y_i^{(n)}$ for all i for $n = 1, 2, ...$, and thus their limits obey $z_i^\infty \geq y_i^\infty$.

For $n = 1$, we get

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^{\gamma} \geq \left(\sum_k v_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^{\gamma} = y_i^{(1)} \tag{4.28}$$

for all i, using that $w_{ik} \geq v_{ik}$ and that the power function x^{γ} is increasing with its argument.

Assuming that $z_i^{(n)} \geq y_i^{(n)}$ for all i for some $n > 0$, then

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} \geq \left(\sum_k v_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} = y_i^{(n+1)} \tag{4.29}$$

using the same properties as before, which proves the assertion for all n by induction.

Remembering that the expected return $V(s_i)$ is increasing with z_i, we conclude that the expected return obtained from policies restricted on the boundaries of the simplices is no better than the original critical value of the expected return.

□

4.5.4 Particular examples

Here we summarize the main results and specialize them to specific cases. We assume $0 < \gamma < 1$, $\alpha > 0$ and $\beta \geq 0$ and use the notation $z_i = \exp(\alpha^{-1}\gamma V^*(s_i))$, where $V^*(s)$ is the optimal expected return, $p_{ijk} = p(s_j|s_i, a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i, a_k)$, where $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a)$.

Action-state entropy maximizers

Agents that seek to maximize the discounted action-state path entropy follow the optimal policy

$$\pi^*(a_k|s_i) = \frac{1}{Z_i} \left(w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \right) \tag{4.30}$$

with

$$Z_i = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk'}} \tag{4.31}$$

The matrix with coefficients $w_{ik} \in \{0, 1\}$ indicate whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$).

The expected return (state-value function) in terms of the z variables obeys

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \tag{4.32}$$

Action-only entropy maximizers

Agents that ought to maximize the time-discounted action path entropy correspond to the above case with $\beta = 0$, and therefore the optimal policy reads as

$$\pi^*(a_k|s_i) = \frac{1}{Z_i}\left(w_{ik} \prod_j z_j^{p_{ijk}}\right) \tag{4.33}$$

with

$$Z_i = \sum_k w_{ik} \prod_j z_j^{p_{ijk}} \tag{4.34}$$

The state-value function in terms of the z variables obeys

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} \prod_j z_j^{p_{ijk}} \tag{4.35}$$

Entropy maximizers in deterministic environments

In a deterministic environment $p_{i,j(i,k),k} = 1$ for successor state $j = j(i,k)$, and zero otherwise. In this case, at every state i we can identify an action k with its successor state j. Therefore, the optimal policy is

$$\pi^*(a_k|s_i) = \frac{w_{ij} z_j}{Z_i} \tag{4.36}$$

with

$$Z_i = \sum_j w_{ij} z_j \tag{4.37}$$

The state-value function in terms of the z variables reads

$$z_i^{\gamma^{-1}} = \sum_j w_{ij} z_j \tag{4.38}$$

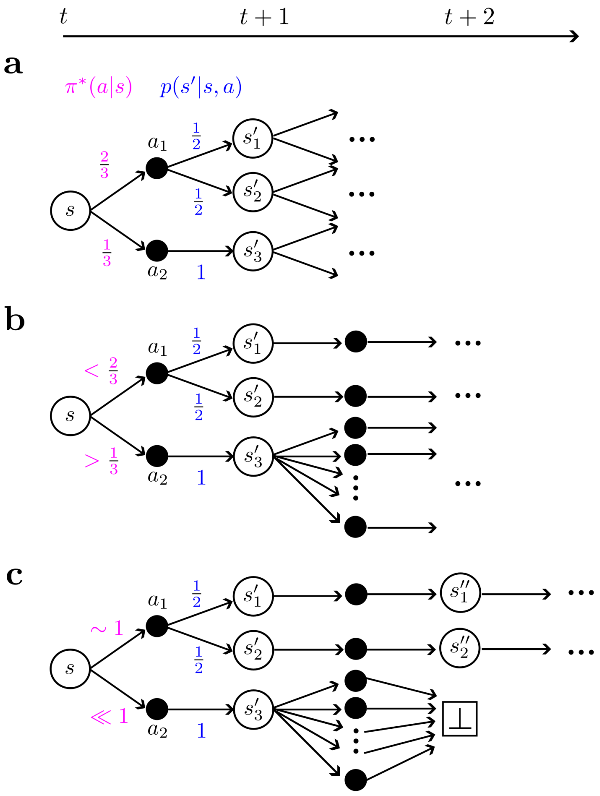


Figure 4.6: H agents determine stochastic policies that maximize occupancy of future action-state paths. In all panels, the three successive dots indicate that the future looks the same for all the states or actions involved from that point onwards. (a) At time t, the agent is faced with determining the optimal policy at state s. Given that taking action a_1 can stochastically lead to two distinct states s'_1 and s'_2, the optimal policy gives action a_1 twice the probability weight than to action a_2 (which only induces a deterministic transition to state s'_3). From time $t+1$, the future looks the same from all three states s'_i. (b) If the future does not look the same, and actually there are many more actions available at state s'_3 compared to s'_1 and s'_2, then more weight should be given to action a_2 than if the future was the same. (c) If, however, all the actions available at state s'_3 lead you to an absorbing state, almost zero weight should be given to action a_2.

4.5.5 Experiments

In this subsection, we present the details for the numerical simulations performed for the different experiments in the manuscript. First, we discuss the construction of the H and R agents, and afterwards we present the details of each particular experiment.

H agent

In all the experiments presented, we introduce the H agent, whose name comes from the usual notation for using H to denote entropy. Therefore, the objective function that this agent maximizes in general is Eq. (4.2). As described in subsection 4.5.4, the α and β parameters control the weights of action and next-state entropies to the objective function, respectively. Unless indicated otherwise, we always use $\alpha = 1, \beta = 0$ for the experiments. It is important to note, as we have done before, that if the environment is deterministic, then the next-state entropy $\mathcal{H}(S'|s,a) = -\sum_{s'} p(s'|s,a) \ln p(s'|s,a) = 0$, and therefore β does not change the optimal policy, Eq. (4.7).

We have implemented the iterative map, Eq. (4.8), to solve for the optimal value, using $z_i^{(0)} = 1$ for all i as initial condition. Theorem (3) ensures that this iterative map finds a unique optimal value regardless of the initial condition in the first orthant. To determine a degree of convergence, we compute the supremum norm between iterations,

$$\delta = \max_i |V_i^{(n+1)} - V_i^{(n)}|,$$

where $V_i = \frac{\alpha}{\gamma} \log(z_i)$, and the iterative map stops when $\delta < 10^{-3}$.

R agent

We also introduce a reward-maximizing agent in the usual RL sense. In this case, the reward is $r = 1$ for living and $r = 0$ when dying. In other words, this agent maximizes life expectancy. Additionally, to emphasize the typical reward-seeking behavior and avoid degenerate cases induced by the tasks, we introduced a small reward for the Four-room grid world (see below). In all other aspects, the modelling of the R agent is identical to the H agent. To allow for reward-maximizing agents to display some stochasticity, we used an ϵ-greedy policy, the best in the family of ϵ-soft policies (Sutton et al., 1998). At any given state, a random admissible action is chosen with probability ϵ, and the action that maximizes the value is chosen with probability $1 - \epsilon$. Given that the world models $p(s'|s,a)$ are known and the environments are static, this ϵ-greedy policy does not serve the purpose of exploration (in the sense of learning), but only to inject behavioral variability.

Therefore, we construct an agent with state-independent variability, whose value function satisfies the optimality Bellman equation for this ϵ-greedy policy,

$$V_\epsilon(s) = (1-\epsilon) \max_a \sum_{s'} p(s'|s,a) \left(r + \gamma V_\epsilon(s')\right) + \frac{\epsilon}{|\mathcal{A}(s)|} \sum_{a,s'} p(s'|s,a) \left(r + \gamma V_\epsilon(s')\right), \tag{4.39}$$

where $|\mathcal{A}(s)|$ is the number of admissible actions at state s. To solve for the optimal value in this Bellman equation, we perform value iteration (Sutton et al., 1998). The ϵ-greedy policy for the R agent is therefore given by

$$\pi(a|s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(s)|}, & \text{if } a = \arg\max_{a'} \sum_{s'} p(s'|s,a') \left(r + \gamma V_\epsilon(s')\right) \\ \frac{\epsilon}{|\mathcal{A}(s)|}, & \text{otherwise} \end{cases}$$

where ties in $\arg\max$ are broken randomly. Note that if $\epsilon = 0$, we obtain the usual greedy optimal policy that maximizes reward.

Four-room grid world

Environment The arena is composed of four rooms, each having size 5×5 locations where the agent can be in. From each room, the agent can go to two adjacent rooms through small openings, each located in the middle of the wall that separates the rooms. At each of these rooms, there is a food source located in the corner furthest from the openings. See Fig. 4.2 for a graphic description. Unless indicated otherwise, the discount factor is set to $\gamma = 0.99$.

States The states are the Cartesian product between (x, y) location and internal state u, which is simply a scalar value between a minimum of 0 and a maximum capacity of 100. All states such that $(x, y, u = 0)$ are absorbing states, independently of the location (x, y). The particular internal state $u = 100$ is the maximum capacity for energy, such that even when at a food source, this internal state does not change. Therefore, the number of states in this experiment is $|\mathcal{S}| = 104\,\text{external states} \times 101\,\text{internal states} = 10504$.

Actions The agent has a maximum of 9 actions: `up, down, left, right, up left, up right, down left, down right`, and `nothing`. Whenever the agent is close to a wall, the number of available actions decreases such that the agent cannot choose to go into walls. Finally, whenever the agent is in an absorbing state, only `nothing` is available.

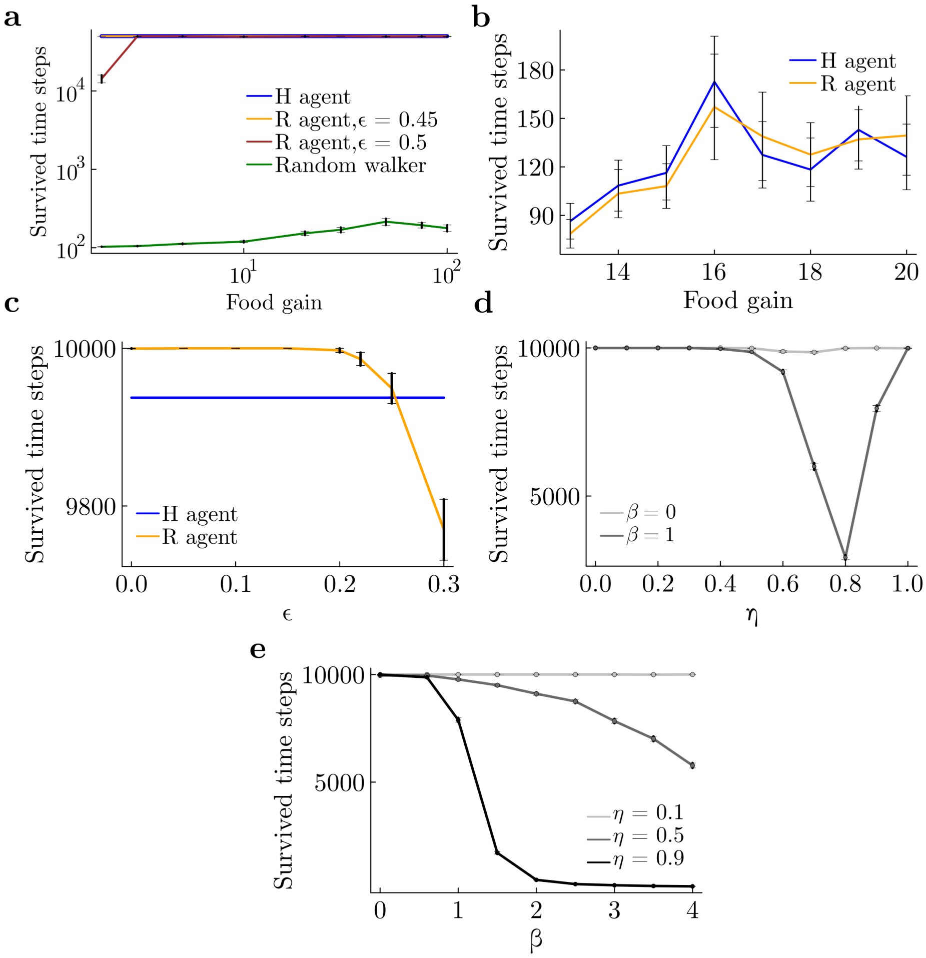


Figure 4.7: Survivability for the experiments considered in our work. (a) Survivability of the various agents tested in the four-room grid world. At each 5E4 timestep episode, we recorded the survived time and averaged across episodes. (b) Survivability of the mouse for both H and R agents. (c) Survivability for the cartpole (Sec. 4.5.5) in the deterministic arena for the H agent and the ϵ-greedy R agents, $\gamma = 0.98$. (d) Survivability for cartpole (Sec. 4.5.5) in the stochastic arena for the $\beta = 0$ and the $\beta = 1$ H agents. $\gamma = 0.99$. (e) Survivability of the cartpole (Sec. 4.5.5) H agents as a function of β, for various values of η. $\gamma = 0.99$

Transitions At any transition, there is a cost of 1 unit of energy for being alive. On the other hand, whenever the agent is located at a food source, there is an increase in energy that we vary parametrically that we call food gain g. For example, if the agent is in location $(2, 1)$ at time t and moves towards $(1, 1)$ (where food is located), the change in energy would be $\Delta u_t = -1$, given that the change in internal energy depends only on the current state and action. If the agent decides to stay in $(1, 1)$ at time $t + 1$, then $\Delta u_{t+1} = -1 + g$.

R agent As stated above, in this experiment we introduced an extra reward for the R agent when it reaches the food source. The magnitude is small compared to the survival reward ($1E - 5$ smaller) and it mainly serves to break the degeneracy of the value function. The variability of the R agent is thus coming purely from the ϵ-greedy action selection.

Survivability To allow for the maximum uniform variability for the R agent, we tested various values for ϵ and observed the survivability of the agents as a function of ϵ, across all the food gains tested (see Results section). The value of ϵ for which the R agent still survives as much as the H agent is $\epsilon = 0.45$ (see Figure 4.7a).

Noisy room In this variation for the experiment, there is a room (the bottom right room) where transitions are uniformly random for all actions, across all possible neighboring locations. That is, for any location s_{nr} in the noisy room, and any a available at that location, given that it has $n(s_{nr})$ total neighbours (including the same location),

$$p(s'|s_{nr}, a) = \begin{cases} \frac{1}{n(s_{nr})} & \text{for } s' \in \text{neighbours} \\ 0 & \text{otherwise} \end{cases}$$

Predator-prey scenario

Here we provide all details of the simulated experiments. Results are shown in Fig. 4.3.

Environment The environment is similar to that one used for the 4-room grid world described in 4.5.5. Apart from the agent (prey), there is also another moving subject (predator) with a simple predefined policy. The grid world consists of a "home" area, a rectangle 2x3 where the agent may enter, but the predator cannot. This home area has a small opening that leads to a bigger 4x7 rectangle arena available for both the agent and the predator. The only food source is located at the bottom-right corner of the common part of the arena, so that the agent needs to leave its home to boost its energy. Additionally, there is an obstacle which separates the arena in two parts with two openings, above and under the obstacle. This obstacle allows the agent to “hide” from the predator behind it.

States The location of the predator is part of the agent’s state, such that a particular state consists of the position of the agent, the position of the predator and the amount of energy of the agent. For this case, we set the maximum amount of energy F equal to the food gain. Positions are 2-dimensional, and therefore the states are 5-dimensional. In the used arena there are 33 possible locations for the agent and 26 ones for the predator, so that the total number of states ranges from 11154 for $F = 13$ to 17160 for $F = 20$.

Actions The agent has the same actions as in the four-room grid world. The maximum number of available actions is therefore 9. Moving towards obstacles or walls is not allowed.

Transitions The agent loses one unit of energy every time step and increases the amount of energy up to a given maximum capacity level F only at the food source. If the position of both the agent and the predator are the same, then the agent is "eaten" and moves to the absorbing state of death as well as in the case of energy equal to 0. After entering the absorbing state the agent stays there forever.

The predator also moves as the agent (horizontally, vertically, diagonally on one step or to stay still). Steps of the agent and the predator happen synchronously. The predator is “attracted” to the agent: the probability of moving to some direction is an increasing function on the cosines $\cos\alpha_k$ of the angle α_k between this direction of motion k and the direction of the radius vector from the predator to the agent. In particular, this probability is

$$p_k^c = C^{-1}\exp(\kappa\cos\alpha_k) \tag{4.40}$$

where κ is the inverse temperature of the predator and $C = \sum_k \exp(\kappa\cos\alpha_k)$ is a normalization factor. These probabilities are computed only for motions available at the current location of the predator, so that e.g. for the location at the wall the motions along the wall are taken into account, but not the motion towards the wall.

Goal The goal of the H agent is to maximize discounted action entropy, and thus to find the optimal state-value function using the iterative map in Eq. (4.8) with $\mathcal{H}_{ik} = 0$ ($\beta = 0$). While using the iterative map, we take advantage of the fact that given an action the physical transition of the agent is deterministic, but the physical transition of the predator is stochastic. Therefore, the sum over successor states j in Eq. (4.8) is simply a sum over the predator successor states.

Parameters $\gamma = 0.98$, $F = 15$ (if another value between 13 and 20 not mentioned), $\kappa = 2$. Simulation time is 5000 steps.

Counting rotations We define a clockwise (counterclockwise) half-rotation as the event when the agent came from the left part of the arena to the right part over the field above (under) the wall and from the right part to the left one over the field under (above) the wall without crossing the vertical line of the wall in between. One full rotation consists of two half-rotations in the same directions performed one after another. We counted the number of full rotations in both directions in 70 episodes of 500 time steps each for both H and R agents for different values of the food gain F. Error bars were computed based on these 70 repetitions. The fraction of clockwise rotations to total rotations (sum of clockwise and anticlockwise rotations) for different values of F is shown at Fig. 4.3.

Survivability The ϵ-greedy R agents display some variability that depends on ϵ. To select this parameter, we matched expected lifetimes (measured in simulations of 5000 steps length) between the H and R agents, separately for every F. Lifetimes are plotted in Figure 4.7b.

Videos We have generated one video for the H agent (Video 2) and another for the R agent (Video 3), both for $F = 15$, $\kappa = 2$, and $\epsilon = 0.06$ for the R agent so as to match their expected lifetimes as described above. In the videos, green vertical bar indicates the amount of energy by the agent at current time. When the agent makes at least one full rotation around the wall, it is indicated by the written phrase “clockwise rotation” or “anticlockwise rotation”. Black vertical arrow indicates direction (‘up’ for clockwise and ‘down’ for anticlockwise directions) of the half-rotation in the part of arena left from the wall.

Cartpole

Environment A cart is placed in a one-dimensional track with boundaries at $|x| = 1.8$. It has a pole attached to it, that rotates like an inverted pendulum with its pivot point on the cart.

States The dynamical system can be described by a four-dimensional external state (x, v, θ, ω), where x is the position of the cart, v is its linear velocity, θ is the angle of the pole with respect to the vertical which grows counterclockwise, and ω is its angular velocity. In this case, we model the internal state u simply with the binary variable `alive, dead`, where the agent enters the absorbing state `dead` if its position exceeds the boundaries, or if its angle exceeds 36 degrees. This amplitude of angles is larger than that typically assumed (12 degrees in (Brockman et al., 2016)), and therefore our system is allowed to be more non-linear and unstable. The state space is $[-1.8, 1.8] \times (-\infty, \infty) \times [-36, 36] \times (-\infty, \infty) \times \{0, 1\}$. To solve for the state value function in Eq. (4.8), we discretize the state space by setting a maximum value for the velocities. Given all the parameters (allowed x and θ, magnitude of the forces, masses of cart and pole, length of pole and gravity, below), we empirically set the maximum values for $|v| = 6$ and $|\omega| = 3$, which the cart actually never exceeds. Therefore, we computed the state value function in a $31 \times 31 \times 31 \times 31 \times 2$ grid (number of states = 1.8×10^6).

Actions Any time the agent is `alive`, it has 5 possible actions: forces of $\{-40, -10, 0, 10, 40\}$, where zero force is understood as `nothing`. If the agent is `dead`, then only `nothing` is allowed.

Transitions This dynamical system is a standard task in reinforcement learning, namely the `cartpole-v0` system of the OpenAI gym (Brockman et al., 2016). The solution of this dynamical system is given in Ref. (Florian, 2007), where we use a frictionless cartpole. The equations for angular and linear accelerations are thus

$$\ddot{\theta} = \frac{-g\sin(\theta) + \frac{\cos(\theta)}{M+m}\left(-F + m\dot{\theta}^2 l \sin(\theta)\right)}{l\left(\frac{4}{3} - \frac{m\cos^2(\theta)}{M+m}\right)} \tag{4.41}$$

$$\ddot{x} = \frac{1}{\cos(\theta)}\left(\frac{4}{3}l\ddot{\theta} - g\sin(\theta)\right). \tag{4.42}$$

Given a force F, a deterministic transition can be computed from these dynamical rules, and a real-valued state transition is observed by the agents.

R agent The reward signal is 1 each time the agent is alive and 0 otherwise. To allow for some variability in the action selection of the R agent, we implement an ϵ-greedy action selection as described above. For exposition purposes, in the manuscript we set $\epsilon = 0.0$, but we also compared to an R agent with ϵ chosen such that average lifetimes between H and R agents are matched (see Fig. 4.7c and Fig. 4.8).

Parameters Mass of the cart $M = 1$, mass of the pole $m = 0.1$, length of the pole $l = 1$, acceleration due to gravity $g = 9.81$, time discretization $\Delta t = 0.02$. Unless specified differently, the discount factor was set to $\gamma = 0.98$.

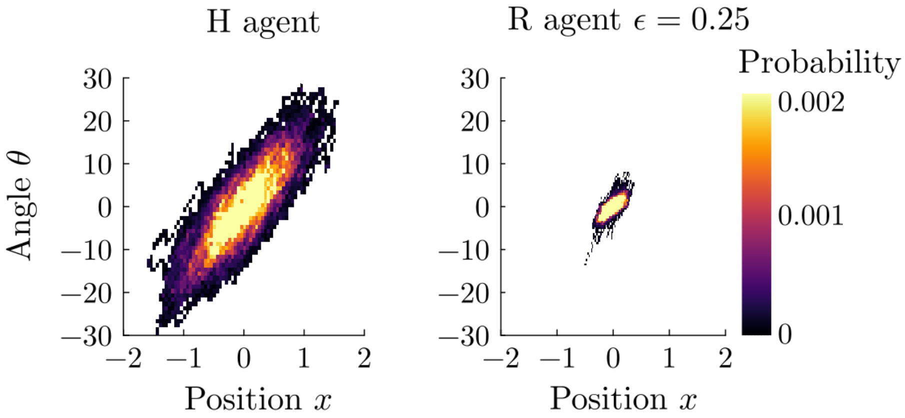


Figure 4.8: Histogram of angles and locations visited for the cartpole, as in Fig. 4.4 of the main manuscript, for the H agent (left) and ϵ-greedy R agent (right), with ϵ chosen such that H and R agents' lifetimes are similar (see Fig. 4.7c).

Value interpolation The observed external state is a continuous four-dimensional variable, so we need to approximate the value function. In order to do so, we simply discretized the state space as described above, and use value iteration as described in Eq. (4.6) in these grid points by performing a linear value interpolation for the successor states at each iteration. During a particular episode, the observed states might not be the same as the ones in the grid, so in order to compute the optimal policy at these states, we perform the same type of value interpolation as in the value iteration stage.

Stochastic arena We introduced a slight variation to the environment, where the $x > 0$ half of the arena is noisy: agents choose an action (force), but the intended state transition of applying such an action fails with probability η and succeeds with probability $1 - \eta$. This is implemented as follows: given any state-action pair (s, a) for which $x > 0$, there are two possible successor states, one corresponding to the intended action (force) chosen, and the other one corresponding to a zero force action:

$$p(s'|s,a) = \begin{cases} 1, & \text{if } x < 0 \text{ and } s' \leftarrow (s,a) \\ 1-\eta, & \text{if } x > 0 \text{ and } s' \leftarrow (s,a) \\ \eta, & \text{if } x > 0 \text{ and } s' \leftarrow (s,0) \end{cases} \tag{4.43}$$

This stochasticity lets us differentiate between action path occupancy maximizers and action-state path occupancy maximizers by choosing any positive real value of β in Eq. (4.1), because $\beta > 0$ agents will have a natural tendency to prefer $x > 0$ locations.

Agent-pet scenario

An agent and a pet move in an arena with degrees of freedom that depend on the actions made by the agent, as explained next in detail.

Environment A 5×5 arena. The middle column of arena can be blocked by a fence, a vertical obstacle that the pet cannot cross. The agent can cross it freely regardless of whether it is open or closed. The agent can open or close the fence by performing the corresponding action when visiting the lever location, at the left bottom corner.

States The system's state consists of the Cartesian product of agent´s location, pet's location and binary state of the fence. So, the number of states is 1250. For the sake of simplicity there is no internal states for the energy, and thus there are not absorbing states. The initial states of the agent and pet at the start of each episode are the middle of the second column and the right lower corner of the arena, respectively.

Actions As in Sec. 4.5.5 the agent's actions are movements to one of the 8 neighbour locations as well as staying on the current one. Additionally, if the agent is on the "lever" location, an additional action is available, namely to open or close the fence, depending on its previous state.

Transitions The pet has same available movements as the agent when the fence is open. The pet performs a random transition to any of the neighbour locations, or stay still, with the same probability. If the agent closes the fence, then the pet can only move on the side where it lies when closed. For simplicity, if the fence is closed by the agent when the pet lies in the middle column, then the pet can only move to the right or left locations such that it will be at one side of the fence in the next time step.

Goal The goal of the H agent is to maximize discounted action-state entropy using the iterative map in Eq. (4.8) with $\alpha = 1$ and $\beta \in [0, 1]$, parameters that measure the weight of action and state entropies, respectively. As in the prey-predator example, we take advantage of the fact that given an action the physical transition of the agent is deterministic, while the physical transition of the pet is stochastic. Thus, the product over successor states j in Eq. (4.8) is a product over the pet successor states.

Simulation details We ran simulations for several values of β, from 0 to 1 in 0.1 steps, to interpolate between pure action entropy ($\beta = 0$) and action-state entropy ($\beta = 1$). We measured the fraction of time the gate was open using episodes of 2000 steps averaged over 70 simulations for each β, shown in Fig 4.5. Heat-maps in that figure correspond to the occupation probability by the pet for $\beta = 0$ (left panel) and $\beta = 1$ (right panel) using an episode of 5000 steps.

4.5.6 Relationship to Maximum Entropy Reinforcement Learning and goal directedness

The objective of maximizing action-state path entropy in Eq. (4.2) for the special case $\beta = 0$ can be obtained from the maximum entropy reinforcement learning (MaxEnt RL) formulation (Todorov, 2009; Ziebart, 2010; Haarnoja et al., 2018)

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right) \Big| s_0 = s \right], \tag{4.44}$$

by setting the reward $r(s, a) = 0$ for all states and actions, and therefore there is no difference between the two approaches in this particular case. However, this reduction obscures the fact that we can generate goal-directed behaviors in H-agents *without* the need of specifying rewards –indeed, this is one of the main accomplishment of our work. To see this, we first quantify how a MaxEnt RL agent gets reward in the four-room grid world defined in Supplemental Sec. 4.5.5, as a function of the temperature parameter α. In this case, a sensible goal is "eating food" (that is, defining $r(s, a) = 1$ at the food locations, and zero everywhere else). Trivially, when $\alpha \ll 1$ in Eq. (4.44), the goal is simply to maximize the future expected reward, equivalent to the ϵ-greedy R agent defined in Supplemental Sec. 4.5.5, for $\epsilon = 0$ (Figure 4.9a, leftmost points). In contrast, for $\alpha \gg 1$, we recover the H agent in practice (due to the environment being deterministic). In this case, the agent mostly focuses on maximizing future expected entropy, and getting small eating rate (Figure 4.9a, rightmost points). Therefore, the temperature α quantifies how "goal directed" the agent should be, where the goal here is understood as getting food, and the entropy term is understood as a regularizer that promotes exploration of the arena.

To aid in showing our central result that an extrinsic reward is not necessary for "goal directed behavior", we take the H agent and vary its energy capacity (see Supplemental Sec. 4.5.5). For large capacities, the H agent can largely ignore the food most of the time, obtaining small eating rate (Figure 4.9b, right-most points). This is because food is conceived as the means to accomplish the goal of maximizing future path occupancy. In contrast, when capacity is small, the H agent needs to get the food much more frequently to avoid the absorbing state, thus

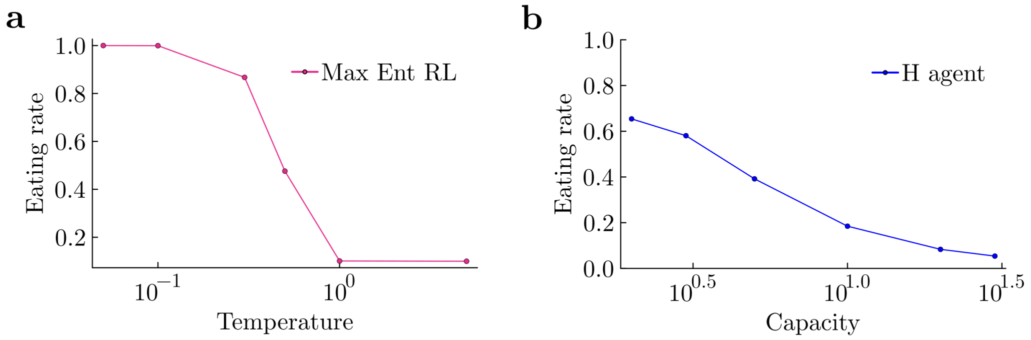


Figure 4.9: Reward is not necessary for "goal-directed" behavior. (a) Eating rate as a function of the temperature parameter α in Equation (4.44) for a MaxEnt RL agent in the four-room grid world. (b) Eating rate as a function of the capacity for an H agent in the four-room grid world.

getting much higher eating rates (Figure 4.9b, leftmost points). The remarkably strong qualitative similarities between the two panels in the figure show that by reinterpreting the concept of reward, one can forego the need of specifying a reward function, and focus on more universal principles of behavior.

4.5.7 Non-additivity of mutual information and channel capacity

Here we show that mutual information over Markov chains does not obey the additive property. It suffices to prove our statement for paths of length two. Thus, we ask whether the mutual information between actions (a_0, a_1) and states (s_1, s_2) given initial state s_0

$$\text{MI}_{\text{global}} = \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0, s_1, a_1, s_2|s_0)}{p(a_0, a_1|s_0)p(s_1, s_2|s_0)}$$

equals the sum of the per-step mutual information

$$\begin{aligned}\text{MI}_{\text{local}} &= \sum_{a_0,s_1} p(a_0, s_1|s_0) \ln \frac{p(a_0, s_1|s_0)}{p(a_0|s_0)p(s_1|s_0)} \\ &+ \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1, s_2|s_1)}{p(a_1|s_1)p(s_2|s_1)}\end{aligned}$$

where $p(a_0, s_1, a_1, s_2|s_0) = \pi(a_0|s_0)p(s_1|s_0, a_0)\pi(a_1|s_1)p(s_2|s_1, a_1)$ and $p(a_0, s_1|s_0) = \pi(a_0|s_0)p(s_1|s_0, a_0)$. Using Bayes' rule and the Markov property, the above quantities can be rewritten as

$$\begin{aligned}
\text{MI}_{\text{global}} &= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0, a_1|s_0, s_1, s_2)}{p(a_0, a_1|s_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{p(a_0, a_1|s_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{\pi(a_0|s_0)p(a_1|s_0, a_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{\pi(a_0|s_0)\sum_s \pi(a_1|s)p(s|s_0, a_0)} \\
&= \sum_{a_0,s_1} p(a_0, s_1|s_0) \ln \frac{p(a_0|s_0, s_1)}{\pi(a_0|s_0)} \\
&+ \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1|s_1, s_2)}{\sum_s \pi(a_1|s)p(s|s_0, a_0)}
\end{aligned}$$

and

$$\text{MI}_{\text{local}} = \sum_{a_0,s_1} p(a_0, s_1|s_0) \ln \frac{p(a_0|s_0, s_1)}{\pi(a_0|s_0)} + \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1|s_1, s_2)}{\pi(a_1|s_1)}$$

The quantities $\text{MI}_{\text{global}}$ and MI_{local} are remarkable similar except for the denominator in the $\ln$ of the last term in each expression. Therefore, equality between $\text{MI}_{\text{global}}$ and MI_{local} holds iff

$$\begin{aligned}
&\sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \sum_s \pi(a_1|s)p(s|s_0, a_0) \\
&\qquad\qquad = \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \pi(a_1|s_1)
\end{aligned}$$

which is not true for all choices of policy and transitions probabilities. To see this, take a Markov chain where the action $a_0 = 0$ from $s_0 = 0$ is deterministic, but results in two possible successor states $s_1 = 1$ or $s_1 = 2$ with equal probability $1/2$. From $s_1 = 1$ the policy takes actions $a_1 = 1$ and $a_1 = 2$ with probability $1/2$. From $s_1 = 2$ the policy is deterministic, that is, $a_1 = 3$ with probability 1. A simple calculation shows that the left side equals $-\frac{3}{2}\ln 2$, while the right side equals a different quantity, $-\frac{1}{2}\ln 2$.

Chapter 5

DISCUSSION

In the previous chapters, we have introduced two main models of decision making: the breadth–depth (BD) dilemma and behavior as the occupation of action-state path space. These works were discussed separately, and in this chapter, we aim to connect them further between each other and with the frameworks presented in the Introduction.

5.1 Summary of contributions

Breadth–depth dilemma

In the BD dilemma, we have explored the influence of *sampling capacity* in strategic decision-making scenarios, where feedback is delayed. As was discussed, this situation not only is relevant for actual real decisions, but it also allows us to study in isolation a thus-far neglected decision-making tradeoff: choosing how to allocate resources to choose between alternatives given fixed sampling restrictions. In Chapter 2, we presented a simplified model of this tradeoff where both the sampling resources and the outcomes of querying the alternatives are discrete (although see Supplementary Fig. 2.7 for the case of Gaussian outcomes). The alternatives are characterized by a real number, the probability of success p_i, with a known, fixed Bernoulli process determining the outcomes of the sampling. We set up the problem of, given a fixed number of samples that we called capacity, finding the best *resource allocation* possible in order to choose the option with the highest probability p_i, given that no reallocation was possible during the actual Bernoulli processes. In other words, we defined a **utility function** to be about the allocation of resources, assuming that the problem was about eventually choosing the *best* option. (Determining *which* option has the highest probability of success given the actual outcomes is a trivial problem, see Sec. 2.4.4.) The rest of the chapter was about characterizing the allocations that maximize this utility function, as a function of the sampling capacity of the agent. The main finding of this work was that, when capacity *grows*, the rate at which the number of alternatives are sampled, relative to capacity, actually *diminishes*, therefore given an increasing emphasis on depth over breadth. This result showcases the value of a normative approach (i.e. searching for the allocation that maximizes the utility) to interpret real choice: it lets us find candidate reasons for observed behavior without *ad hoc* assumptions. In particular, it lets us reinterpret the observation that humans choose to focus deeply on a small subset of options, instead of broadly, which might actually reflect a high sampling capacity (see the Discussion of Chapters 2 and 3 for more details).

In Chapter 3, we presented a model of the same BD tradeoff where both the sampling resources and the outcomes of querying the alternative are continuous. Alternatives are also characterized by a real number, the mean μ_i, and in this case a drift–diffusion process determines the outcome of the sampling process, given μ_i for each alternative and diffusion noise σ. For this case, the capacity of the agent to sample the environment was defined as the ratio between the precision of the observations and the precision of the distribution that characterizes the richness of the environment, a measure of discriminability of the quality of the options. Therefore, the sampling capacity of the decision maker takes the environment statistics into account, thus contextualizing the decision maker's resources and characterizing how well they are suited to solve a given task in the world. Allocating *capacity* opened up a duality: we can think that agents either allocate sampling *time*, or they allocate sampling *precision*. This is an important distinction, as the former possibility is typically associated with a single agent in a sequential sampling scenario, whereas the former can be associated to a distributed system of samplers, each with an allocated precision (see Fig. 3.2). One of the main results of this work was to find that, for small capacity and for any environment, the optimal number of options to sample is fixed and equal to five (provided there are at least five options available). We provided the proof for this limit, which agrees with empirical simulation tests. For large capacity, the optimal number of options to sample grows *sublinearly* for a variety of environments, a fact that matches the result for discrete resources, showing a sign that, for increasing capacity, there is an increasing emphasis of depth over breadth in the optimal tradeoff.

The models of the breadth–depth dilemma make certain assumptions that can be analyzed with the frameworks presented in the Introduction. First off, we assume that the objective is to find the best resource allocation given a particular capacity. Making use of Tinbergen's current utility question, we suggest that finding the best option amongst many is a fundamental cognitive ability to the survival of the decision maker. In other words, we propose **why** certain allocations will be preferred to others, given the agent's **constraints**. Secondly, the problem assumes familiarity with the environment, in the sense that the overall distribution of richness in the repertoire of alternatives is known. Therefore, we do not make use of

Tinbergen's development question, as we do not inquire **how** real agents can learn the parameters of this distribution. Thirdly, we assume a specific sampling process, in which feedback is only available after the deployment of the resources. It is a constraint at the level of Tinbergen's mechanistic question, as it directly restricts **how** the sampling should happen, e.g. a distributed system where reallocation of resources is impossible. Finally, the capacity of the agent is thought to be a relatively fixed parameter that characterizes its **constraints**, perhaps brought by evolution, that answer **why** an agent's allocations should follow a particular rule.

On the other hand, there is a hierarchy of processes in the breadth–depth dilemma. The actual process of sampling and choosing the best option is a "low level" process; the situation in which the agent is assumed to actually be. The higher level process, then, is the allocation of resources. Its optimal solution need not be computed in real time, nor "consciously" by the agent. These two levels can be read from the two maximizer operators in Eqs. 2.7 and 3.5. Each level of the hierarchy can be inspected from Marr's levels of analysis. In terms of the low level, sampling process, the BD tradeoff assumes a particular algorithm, which is influenced by the capacity of the agent. In this sense, the constraints that enter the algorithmic level, directly influence the computational level – choosing the best allocation of samples. By building the higher level of the hierarchy, we managed to transform the problem into a metareasoning problem, thus providing a computational theory at this level. There are no assumptions about *how* the optimal allocations are found. This critical distinction is aligned with the idea of bounded optimality, in which agents are optimized at *design time*, instead of at *run time*, therefore escaping the problem of perfect rationality (and perfect metarationality, and so forth). The actual mapping of this problem to the bounded optimality framework, in the style of ecological bounded optimality (Lewis et al., 2014) is a task for the future. By explicitly associating the features of the BD dilemma with Tinbergen's four questions and Marr's levels, we are immediately pointed towards two possible avenues of future research – namely how learning impacts the allocation of resources and how optimal allocations can be found.

Path occupancy maximization

In Chapter 5, we moved away from the problem of optimally allocating resources in the breadth–depth tradeoff in order to maximize utility, to actually asking what utility means for an agent. In order to do so, we needed to introduce the concept of intrinsic motivation, the idea that agents have an internal notion of what they want to achieve, without the need to externally craft a reward function. We proposed a principle that is inspired by the variability of natural behavior by proposing variability itself as the goal. We formalized this idea by formulating the objective as maximizing the occupancy of future paths. We showed that entropy is the only measure consistent with intuitive *desiderata* about path occupancy, and showed that this principle generates complex behavior in various simulated environments.

A crucial ingredient in this framework is that agents do not "choose" actions. Instead, they determine, consciously or unconsciously, the probability distribution over actions, such that the actual action taken is randomly drawn from this distribution. While agents can still take actions deterministically (by setting the probability of a particular action equal to one), they are discouraged to do so by seeking variation in actions and states, since this leads to a variation of paths, thus fulfilling the principle. The fact that agents are attracted to regions of state space where both the policy and the transition probabilities are as uniform *as possible* lets us interpret that these agents seek to "maximize agency" (always taking deterministic actions and observing deterministic transitions is "boring" and takes away their agency). The POM principle leads then to *seemingly* goal-directed, yet highly variable, behavior: agents seek states (e.g. food sources) and actions (e.g. moving away from a predator) that allow them to keep occupying future paths in the long run. Seeking those states and actions, from the limited perspective of an external observer, might appear to be reward-seeking behaviors, even if from the internal perspective of the agents, they are only "fuel" to keep going.

At first, the identification of reward as means and not as goals might seem simply as a semantic problem. In other words, it might be argued that the *reward* in the reward function can be intrinsically generated, such as a homeostatic signal. In fact, Keramati and Gutkin (2014) have introduced an algorithmic model to make standard reinforcement learning compatible with homeostatic regulation by crafting the reward function as the difference in distance to a vector of set points, a model called homeostatic regulated reinforcement learning (HRRL). In this way, if a particular external state and action were associated to particular internal states being closer to their set points (e.g. in hypothalamus), then this would translate as a high reward that can be used to compute value (e.g. in cortico-basal ganglia circuits). Path occupancy maximization agrees in some points to HRRL, in particular to the idea that rewards *should* be represented as internal states (or transitions between states) of the system in question. The main difference, however, is that HRRL assumes the existence of a family of set points (perhaps brought by the phylogeny of the agent in question, or by developing and learning in a particular environment (Juechems and Summerfield, 2019)), but it does not say **why** they exist, **why** they are set to a particular value, or **why** they are relevant for the agent. In contrast, POM makes no assumptions about the internal structure of states that limits the interpretation of intrinsic reward, other than the possible existence of death states, which by construction immobilize the agent in state and actions. The advantage is that desired ranges for internal states emerge from the need to occupy path space, and only those that serve this purpose will be relevant, thus providing candidates for a more complete causal understanding of behavior.

In essence, POM is a **computational theory** of behavior, whereas HRRL uses the reward maximization hypothesis as the computational theory to then construct an **algorithmic theory** of homeostatic regulation by reinforcers.

One of the main results of this work is then to reconceptualize reward as means to occupy path space, instead of the goal. This important distinction also allows us to interpret laboratory settings differently. In laboratory tasks, rewards are usually identified *a priori* (e.g. water for a water-restricted animal), and scientists typically assume the ‘reward hypothesis’ to hold, which says that agents will tend to maximize reward (or reward-rate) (Juechems and Summerfield, 2019). Under both the POM and the HRRL view, this will only be true as long as the agent *needs* the particular reward. Unlike the HRRL, the POM principle can tell you **why** the agent needs the reward, which includes, but is not limited to, its own survival. Hence, this work is best understood as trying to answer Tinbergen’s **why** questions: current utility and evolution, by providing a principle that can be applied to both. While the POM principle is agnostic to **how** behaviors come to be, the work presented here suggests certain mechanisms by which the POM principle is implemented in natural agents – in particular, the notion that rewards are synonymous of transitions between internal states, which allows a natural interpretation of path space occupancy.

Finally, the POM principle relies on the correct identification of **constraints** in the agent-environment loop. As was stated in Chapter 5, in any Markov decision process with full symmetry in the state-action dynamics, a random walk maximizes path space occupancy. However, we start observing more interesting behavior when we introduce constraints that break symmetries, such as energetic limitations, presence of a predator, or a cart and pole dynamical system with specific dynamics. The POM principle then, by construction, produces the most variable behavior that constraints allow.

5.2 Future directions

There are several potential avenues of research to expand the work introduced in this thesis. While this has already been discussed separately, we can delineate more in detail the possible contributions in the future stemming from our work.

Breadth–depth dilemma

The breadth–depth dilemma was introduced in a general setting in Chapter 2, and at the same time it was modeled in two different ways (Chapter 2 and 3). Ultimately, the optimal tradeoff between breadth and depth search will depend on the actual scenario involved. Other scenarios have been developed specifically for the formulation presented in this thesis. For example, a decision tree search under limited resources displays the breadth–depth tradeoff that encourages yet again depth over breadth search for a wide range of parameter values (Mastrogiuseppe and Moreno-Bote, 2022).

As mentioned before, a crucial ingredient for breadth–depth problems is the identification of resource allocation for subsequent sampling. In this thesis, we have delayed feedback about the quality of the options completely in order to study the tradeoff in isolation. However, real decisions combine a plethora of cognitive tradeoffs due to particular constraints in the problem formulation. For example, free-response sequential decisions display a speed–accuracy tradeoff due to the limitation of sequential sampling in noisy environments. The exploration–exploitation (EE) tradeoff appears when a sequential search needs to be performed in uncertain environments in order to maximize future reward. In general, functional constraints necessarily imply tradeoffs that can be of different nature (Del Giudice and Crespi, 2018).

A possibility to expand the research into the BD tradeoff is then to incorporate other constraints into the problem formulation to make it interact with other cognitive trade-offs. For example, if reallocation of resources is allowed at any moment during the sampling phase (perhaps at some cost), then a full fledged interaction between the speed–accuracy tradeoff and the breadth–depth tradeoff can be studied. In fact, there is currently work to explore dynamic allocation of attentional resources by a recurrent neural network when feedback about the initial consideration set is available over time (Damiani and Moreno-Bote, 2023). By stopping evidence accumulation about certain alternatives during the sampling process, the network can obtain a better accuracy about the remaining alternatives, thus improving the reward obtained, and inducing a full-fledged interaction between a depth–breadth tradeoff (induced by a decision about the initial consideration set) and the speed–accuracy tradeoff (by reducing the alternatives considered to improve accuracy, or stopping the accumulation altogether to favor speed). There are two main differences about this approach with respect to the dynamic allocations presented in Chapter 2. First, our dynamic allocations were *static*, in the sense that the number of alternatives to drop per wave were a fixed strategy, whereas in the RNN scenario, the actual observations received by the network drive it to drop certain alternatives. Secondly, the dynamic allocation in Chapter 2 considered a discrete number of waves, whereas the RNN works in continuous time, and accumulates evidence in a similar fashion to the setting in Chapter 3.

Another interesting alternative is to remove the assumption of familiarity with the environment to induce a form of exploration–exploitation tradeoff: allocation of resources needs to happen not only to discover a good option (exploitation), but also to learn the environment statistics (exploration). The decision maker is therefore prompted to plan the resource allocation in a way that they receive information about the environment, which is crucial to discover the optimal allocations in the future. In Chapter 3, we provided a definition of capacity that depends on the discriminability of options $1/\sigma_0^2$, where σ_0^2 is the variance of the prior distribution, as $C = \frac{\sigma_0}{\sigma}T$. If the decision maker *does not* know σ_0 in advance, it is as if they do not know their *true* capacity, which they need to discover. The optimal strategy to allocate resources in order to maximize reward in the long run will be a crucial answer for this BD–EE interaction.

Finally, it is worth mentioning that Vidal et al. (2022) have tested the predictions from Chapter 2 about the optimal tradeoff in an experimental setting with humans that isolates the BD dilemma. They show that humans are capable of trading off breadth versus depth as their sampling capacity is experimentally changed in a way that is consistent with the optimal strategies presented here. Even if there are slight deviations from the prescribed optimal strategies in Chapter 2, humans do a remarkable job at recognizing how their strategies should change once they are able to gather more information about the world. Since their experimental paradigm uses the same Bernoulli process described in Chapter 2, another possibility to explore in the future is to test the predictions of continuous resource allocation (such as time, or precision), either in full isolation as presented in Chapter 3, or the alternatives listed above.

Path occupancy maximization

There are various way to expand the path occupancy maximization principle for the understanding of behavior. As was mentioned in the Discussion of Chapter 4, one of the clearest avenues is to formalize learning for this principle. Because there is no need to learn any reward structure, in this context it is only needed to learn the world model (the transition probability between states given actions). While this is quite a hard problem, there are techniques in the literature that deal with this problem and it is thus not necessarily an issue of this theory (for examples see (Bellemare et al., 2016; Tang et al., 2017)). By necessarily injecting a prior into the learning agent, it would be possible to model optimism (or pessimism) as in standard reinforcement learning by setting the initial state values higher (or lower) than the bounds of the value function (Sutton et al., 1998). In this case, however, optimism is not about the amount of *reward* that a state-action pair can lead to in the future, but about the future path occupancy. In other words, POM agents that are optimistic have the notion that any state-action pair has high path availability, which maps intuitively to the notion of curiosity by encouraging agents to try different courses of action.

In the same line, a possibly important avenue for the POM principle is to use it to discover and exploit skills. One of the main assumptions in this theory is that states and actions are provided, over which paths are constructed. However, there are two issues of different nature with this assumption. First, not only is this not always the case, sometimes the state and action space is so large that learning from path occupancy might not be efficient in covering action-state space, such that agents are trapped in a small region of path space (Campos et al., 2020). This is not a problem by itself given that the objective is *not* covering action-state space, but in practice it might deliver agents that are too limited in what they do and what they can visit. This leads to the second issue, which is that it can be argued that natural agents show much smoother behaviors than those apparent in our results. Using realistic dynamics where actions are analogous or equal to forces make the dynamics smoother, as in the cartpole experiment (Fig. 4.4). However, actions (in this case, forces) *do* vary significantly with a small timescale; a testable, but dubious, consequence from using primitive actions to construct the POM principle. One potential solution is to apply the POM principle to state-skill paths, or state-option paths, where skills are latent variables on which policies can be conditioned to deliver higher-level actions (Eysenbach et al., 2018; Sharma et al., 2020), and options are temporally extended actions that call on primitive actions (e.g. 'getting up from a chair' calls on muscle contractions to deliver the desired higher level action) (Sutton et al., 1999). This inevitably leads to asking what the right level of coarse-graining is for path occupancy maximization, both in space and actions, a question that is undoubtedly critical for agents that make decisions in real time and under resource constraints (Harb et al., 2017). Certainly, when there are no resource constraints, another important direction for POM is to make it scalable – devise ways to deal with large state and action spaces, as well as continuous state and action spaces. Bringing it to the state of the art of machine learning research would mean dealing with sophisticated value function approximators and efficient learning algorithms. While in principle there are no theoretical limitations to do this, this direction is out of the scope of this thesis.

Finally, path occupancy maximization itself need not be the *ultimate* intrinsic motivation approach. In fact, based solely on occupation of state and action spaces, there exists already approaches that deal with a coverage problem: what is the optimal policy to *explore* state space in an efficient way? (See (Hazan et al., 2019; Neu et al., 2017; Liu and Abbeel, 2021; Mutti et al., 2021; Amin et al., 2021).) While the inspiration is radically different, the methods might be similar. In fact, one could pose the problem introduced by POM not as occupying *path* space, but to actually occupy state and action space in a reward-free manner. This objective function implies the existence of a stationary state distribution induced by the optimal policy, by which the agent is pushed to maximize the state or action entropies. We have unified some of these approaches in a general theory entropy regularization in MDPs, where the state and action entropies can have arbitrary weights (Grytskyy et al., 2023), but its implications for learning are still a work in progress. In the end, one of the main inspirations of the POM principle was behavioral variability itself, and thus a natural direction for this work to expand would be to try to understand animal behavior as maximizing path occupancy. As said earlier, the right level of path abstraction will need to be either controlled or inferred. All in all, the POM principle could also inspire a new neuroscience approach to behavior that shies away from the reward hypothesis, bringing animal behavior to more ecological tasks in which the analysis and interpretation of behavior are not contingent to an external reward function set by the experiment design.

Bibliography

Achiam, J. and Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

Adolph, K. E. and Berger, S. E. (2007). Motor development. *Handbook of child psychology*, 2. Publisher: Wiley Online Library.

Amin, S., Gomrokchi, M., Satija, H., van Hoof, H., and Precup, D. (2021). A Survey of Exploration Methods in Reinforcement Learning. *arXiv preprint arXiv:2109.00157*.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press. Google-Books-ID: T5JBLb1cNUgC.

Anderson, P. W. (1972). More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396. Publisher: American Association for the Advancement of Science.

Andrews, G. E. (1998). *The Theory of Partitions*. Cambridge University Press. Google-Books-ID: Sp7z9sK7RNkC.

Asafa, T., Afonja, T., Olaniyan, E., and Alade, H. (2018). Development of a vacuum cleaner robot. *Alexandria engineering journal*, 57(4):2911–2920. Publisher: Elsevier.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3):183. Publisher: American Psychological Association.

Aubret, A., Matignon, L., and Hassas, S. (2022). An information-theoretic perspective on intrinsic motivation in reinforcement learning: a survey. *arXiv preprint arXiv:2209.08890*.

Averbeck, B. B. (2015). Theory of choice in bandit, information sampling and foraging tasks. *PLoS computational biology*, 11(3):e1004164. Publisher: Public Library of Science San Francisco, CA USA.

Awh, E., Barton, B., and Vogel, E. (2007). Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. *Psychological Science*, 18:622–628 10 1111 1467–9280 2007 01949.

Balasubramani, P., Moreno-Bote, R., and Hayden, B. (2018). Using a Simple Neural Network to Delineate Some Principles of Distributed Economic Choice. *Front. Comput. Neurosci*, 12:22.

Ballesta, S. and Padoa-Schioppa, C. (2019). Economic decisions through circuit inhibition. *Current Biology*, 29(22):3814–3824. Publisher: Elsevier.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394. Publisher: SAGE Publications Sage UK: London, England.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846. Publisher: IEEE.

Bates, C., Lerch, R., Sims, C., and Jacobs, R. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, 19:11.

Bateson, P. and Laland, K. N. (2013). Tinbergen's four questions: an appreciation and an update. *Trends in ecology & evolution*, 28(12):712–718. Publisher: Elsevier.

Beach, L. R. (1993). Broadening the definition of decision making: The role of prechoice screening of options. *Psychological Science*, 4(4):215–220. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Bechhofer, R. E. and Kulkarni, R. V. (1984). Closed sequential procedures for selecting the multinomial events which have the largest probabilities. *Communications in Statistics-Theory and Methods*, 13(24):2997–3031. Publisher: Taylor & Francis.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152. Publisher: Elsevier.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Berger, J. O. and Berger, J. O. (1985). *Bayesian analysis*. Springer.

Bettman, J., Luce, M., and Payne, J. (1998). Constructive Consumer Choice Processes. *Journal of Consumer Research*, 25:187–217.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Brady, T. F., Konkle, T., and Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, 11(5):4–4. Publisher: The Association for Research in Vision and Ophthalmology.

Brandstätter, E., Gigerenzer, G., and Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113:409–432 10 1037 0033–295 113 2 409.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2019). Large-Scale Study of Curiosity-Driven Learning. In *ICLR*.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2018). Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Busemeyer, J., Gluth, S., Rieskamp, J., and Turner, B. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, 23:251–263.

Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432. Publisher: American Psychological Association.

Callaway, F., Rangel, A., and Griffiths, T. L. (2021). Fixation Patterns in Simple Choice Reflect Optimal Information Sampling. *PLOS Computational Biology*, 17(3).

Campos, V., Trott, A., Xiong, C., Socher, R., Giro-I-Nieto, X., and Torres, J. (2020). Explore, Discover and Learn: Unsupervised Discovery of State-Covering Skills. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1317–1327. PMLR. ISSN: 2640-3498.

Cazettes, F., Murakami, M., Renart, A., and Mainen, Z. F. (2021). Reservoir of decision strategies in the mouse brain. *bioRxiv*. Publisher: Cold Spring Harbor Laboratory.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016). Combinatorial Multi-Armed Bandit with General Reward Functions. In *Advances in Neural Information Processing Systems*, volume 29, pages 1659–1667.

Churchland, A. K., Kiani, R., and Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702. Publisher: Nature Publishing Group.

Cisek, P. and Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, 33:269–298. Publisher: Annual Reviews.

Cohen, J., McClure, S., and Yu, A. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362:933–942.

Corver, A., Wilkerson, N., Miller, J., and Gordus, A. (2021). Distinct movement patterns generate stages of spider web building. *Current Biology*, 31(22):4983–4997. Publisher: Elsevier.

Costa, V., Mitz, A., and Averbeck, B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron*, 103:533–545 5.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., and Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1):42–100. Publisher: Elsevier.

Cuthill, I. (2005). The study of function in behavioural ecology. *Animal Biology*, 55(4):399–417. Publisher: Brill.

Dagenais, P., Hensman, S., Haechler, V., and Milinkovitch, M. C. (2021). Elephants evolved strategies reducing the biomechanical complexity of their trunk. *Current Biology*, 31(21):4727–4737. Publisher: Elsevier.

Damiani, F. and Moreno-Bote, R. (2023). Optimal dynamic allocation of finite resources for many alternatives decision-making.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., and Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879.

De Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.

Del Giudice, M. and Crespi, B. J. (2018). Basic functional trade-offs in cognition: An integrative framework. *Cognition*, 179:56–70. Publisher: Elsevier.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953. Publisher: Oxford University Press.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, 18:193–222.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70(1):80. Publisher: American Psychological Association.

Diehl, K. and Poynor, C. (2010). Great Expectations?! Assortment Size, Expectations, and Satisfaction. *Journal of Marketing Research*, 47:312–322.

Dietrich, A. (2004). The cognitive neuroscience of creativity. *Psychonomic bulletin & review*, 11(6):1011–1026. Publisher: Springer.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081. Publisher: Elsevier.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, 32(11):3612–3628. Publisher: Soc Neuroscience.

Ebitz, R., Albarran, E., and Moore, T. (2018). Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron*, 97:450–461 9.

Eisenreich, B., Akaishi, R., and Hayden, B. (2017). Control without Controllers: Toward a Distributed Neuroscience of Executive Control. *Journal of Cognitive Neuroscience*, 29:1684–1698.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Eysenbach, B. and Levine, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*.

Feng, S. F., Schwemmer, M., Gershman, S. J., and Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, 14(1):129–146. Publisher: Springer.

Fischer, R. and Plessow, F. (2015). Efficient multitasking: parallel versus serial processing of multiple tasks. *Frontiers in Psychology*, 6:1366. Publisher: Frontiers.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Florian, R. V. (2007). Correct equations for the dynamics of the cart-pole system. *Center for Cognitive and Neural Studies (Coneural), Romania*. Publisher: Citeseer.

Fountas, Z., Sajid, N., Mediano, P., and Friston, K. (2020). Deep active inference agents using Monte-Carlo methods. *Advances in neural information processing systems*, 33:11662–11675.

Fox, R., Pakman, A., and Tishby, N. (2015). Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*.

Fudenberg, D., Strack, P., and Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–84.

Galashov, A., Jayakumar, S. M., Hasenclever, L., Tirumala, D., Schwarz, J., Desjardins, G., Czarnecki, W. M., Teh, Y. W., Pascanu, R., and Heess, N. (2019). Information asymmetry in KL-regularized RL. *arXiv preprint arXiv:1905.01240*.

Gershman, S. J., Horvitz, E., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278. MAG ID: 1892018222.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62:451–482. Publisher: Annual Reviews.

Gigerenzer, G. and Selten, R. (2000). Bounded rationality: The adaptive toolbox. *International Journal of Psychology*, 35:203–204. MAG ID: 1989020845.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Glöckner, A. and Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of experimental psychology: Learning, memory, and cognition*, 34(5):1055. Publisher: American Psychological Association.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593. Publisher: Elsevier.

Gould, S. J. and Vrba, E. S. (1982). Exaptation—a Missing Term in the Science of Form. *Paleobiology*, 8(1):4–15.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, 7(2):217–229. MAG ID: 2141467654 S2ID: 50485a11fc03e14031b08960370358c26553d7e5.

Grytskyy, D., Ramírez-Ruiz, J., and Moreno-Bote, R. (2023). A general Markov decision process formalism for action-state entropy-regularized reward maximization. arXiv:2302.01098 [cs].

Gupta, S. and Liang, T. (1989). Selecting the best binomial population: parametric empirical Bayes approach. *Journal of Statistical Planning and Inference*, 23:21–31 10 1016 0378–3758 89 90036–0.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., and Dragan, A. (2017). Inverse reward design. *Advances in neural information processing systems*, 30.

Hafner, D., Ortega, P. A., Ba, J., Parr, T., Friston, K., and Heess, N. (2020). Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*.

Halpert, H. (1958). Folklore: Breadth versus Depth. *The Journal of American Folklore*, 71:97 10 2307 537679.

Harb, J., Bacon, P.-L., Klissarov, M., Doina Precup, Doina Precup, Doina Precup, and Precup, D. (2017). When Waiting is not an Option : Learning Options with a Deliberation Cost. *arXiv: Artificial Intelligence*. ARXIV_ID: 1709.04571 MAG ID: 2754203286 S2ID: 96e81cabed55630f2ad3e1346300bd7a7a17f060.

Hauser, J. R. and Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research*, 16(4):393–408. Publisher: The University of Chicago Press.

Hausman, K., Springenberg, J. T., Wang, Z., Heess, N., and Riedmiller, M. (2018). Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*.

Hayden, B. Y. and Moreno-Bote, R. (2018). A neuronal theory of sequential economic choice. *Brain and Neuroscience Advances*, 2:2398212818766675. Publisher: SAGE Publications Sage UK: London, England.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR.

Hick, W. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, 4:11–26.

Hills, T., Jones, M., and Todd, P. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119:431–440.

Hills, T., Todd, P., and Goldstone, R. (2010). The central executive as a search process: Priming exploration and exploitation across domains. *Journal of Experimental Psychology: General*, 139:590–609.

Hladky, V. and Havlicek, J. (2013). Was Tinbergen an Aristotelian? Comparison of Tinbergen's four whys and Aristotle's four causes. *Human Ethology Bulletin*, 28(4):3–11.

Horowitz, E. and Sahni, S. (1978). *Fundamentals of computer algorithms*. Computer Science Press.

Horvitz, E. (1987). Reasoning about beliefs and actions under computational resource constraints. *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence*, pages 429–447. MAG ID: 1525110644.

Huxley, J. S. (2009). *Evolution, The Definitive Edition: The Modern Synthesis*. MIT Press. Google-Books-ID: FxtFAQAAIAAJ.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995. Publisher: American Psychological Association.

Jang, A. I., Sharma, R., and Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, 10:e63436. Publisher: eLife Sciences Publications Limited.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR.

Juarrero, A. (2000). Dynamics in action: Intentional behavior as a complex system. *Emergence*, 2(2):24–57. Publisher: Taylor & Francis.

Juechems, K. and Summerfield, C. (2019). Where Does Value Come From? *Trends in Cognitive Sciences*, 23(10):836–850.

Jung, T., Polani, D., and Stone, P. (2011). Empowerment for continuous agent—environment systems. *Adaptive Behavior*, 19(1):16–39. Publisher: SAGE Publications Sage UK: London, England.

Kahneman, D. and Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Keramati, M. and Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811. Publisher: eLife Sciences Publications, Ltd.

Kidd, C. and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460. Publisher: Elsevier.

Kline, S. J. and Rosenberg, N. (2010). An overview of innovation. *Studies on science and the innovation process: Selected works of Nathan Rosenberg*, pages 173–203. Publisher: World Scientific.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE.

Koechlin, E. and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11:229–235.

Korf, R. E. (1985). Depth-first iterative-deepening: An optimal admissible tree search. *Artificial Intelligence*, 27(1):97–109. Publisher: Elsevier.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298. Publisher: Nature Publishing Group.

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857. Publisher: National Acad Sciences.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, 93(3):480–490.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. (2019). Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223. Publisher: MIT Press.

Leibfried, F., Pascual-Diaz, S., and Grau-Moya, J. (2019). A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32.

Levin, I. P., Jasper, J., and Forbes, W. S. (1998). Choosing versus rejecting options at different stages of decision making. *Journal of Behavioral Decision Making*, 11(3):193–210. Publisher: Wiley Online Library.

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization. *Topics in Cognitive Science*, 6(2):279–311. MAG ID: 2121813011.

Lim, S.-L., O'Doherty, J. P., and Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *Journal of Neuroscience*, 31(37):13214–13223. Publisher: Soc Neuroscience.

Liu, H. and Abbeel, P. (2021). Behavior from the void: Unsupervised active pretraining. *Advances in Neural Information Processing Systems*, 34:18459–18473.

Luck, S. J. and Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, 17(8):391–400. Publisher: Elsevier.

Ma, W. J., Husain, M., and Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3):347. Publisher: Nature Publishing Group.

MacNeilage, P. F. and Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, 288(5465):527–531. Publisher: American Association for the Advancement of Science.

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press. Google-Books-ID: Wr34DwAAQBAJ.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry.

Mastrogiuseppe, C. and Moreno-Bote, R. (2022). Deep imagination is a close to optimal policy for planning in large decision trees under limited resources. *Scientific reports*, 12(1):10411. Publisher: Nature Publishing Group UK London.

McNamara, J. M. and Houston, A. I. (1986). The common currency for behavioral decisions. *The American Naturalist*, 127(3):358–378. Publisher: University of Chicago Press.

Mehta, N., Rajiv, S., and Srinivasan, K. (2003). Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Science*, 22(1):58–84. Publisher: INFORMS.

Messick, D. M. (1993). Equality as a decision heuristic. *Psychological Perspectives on Justice: Theory and Applications*, pages 11–31.

Meyer, D. E. and Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological Review*, 104(1):3. Publisher: American Psychological Association.

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. In *Proceedings of the Human Factors Society Annual Meeting*, volume 25, pages 296–300. SAGE Publications Sage CA: Los Angeles, CA.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81. Publisher: American Psychological Association.

Mochol, G., Kiani, R., and Moreno-Bote, R. (2021). Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior. *Current Biology*, 31(6):1234–1244. Publisher: Elsevier.

Mohamed, S. and Jimenez Rezende, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28.

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Computation*, 22(7):1786–1811. Publisher: MIT Press.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496. Publisher: National Acad Sciences.

Moreno-Bote, R., Ramírez-Ruiz, J., Drugowitsch, J., and Hayden, B. Y. (2020). Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences*, 117(33):19799–19808. Publisher: National Acad Sciences.

Moreno-Bote, R., Rinzel, J., and Rubin, N. (2007). Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of Neurophysiology*, 98(3):1125–1139. Publisher: American Physiological Society.

Morgan, P. and Manning, R. (1985). Optimal Search. *Econometrica*, 53:923 10 2307 1912661.

Mutti, M., Pratissoli, L., and Restelli, M. (2021). Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9028–9036. Issue: 10.

Mutti, M. and Restelli, M. (2020). An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5232–5239. Issue: 04.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30.

Neu, G., Jonsson, A., and Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*.

Newton, I. (1675). Isaac Newton letter to Robert Hooke.

Norman, D. A. and Shallice, T. (1986). Attention to action. In *Consciousness and self-regulation*, pages 1–18. Springer.

Olshavsky, R. W. (1979). Task complexity and contingent processing in decision making: A replication and extension. *Organizational Behavior and Human Performance*, 24(3):300–316. Publisher: Elsevier.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286. Publisher: IEEE.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, 16(2):366–387. Publisher: Elsevier.

Petitto, L. A. and Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, 251(5000):1493–1496. Publisher: American Association for the Advancement of Science.

Pratt, S., Mallon, E., Sumpter, D., and Franks, N. (2002). Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant Leptothorax albipennis. *Behavioral Ecology and Sociobiology*, 52:117–127 10 1007 00265–002–0487–.

Ramírez-Ruiz, J., Grytskyy, D., and Moreno-Bote, R. (2022). Seeking entropy: complex behavior from intrinsic motivation to occupy action-state path space. arXiv:2205.10316 [cs, q-bio].

Ramírez-Ruiz, J. and Moreno-Bote, R. (2022). Optimal allocation of finite sampling capacity in accumulator models of multialternative decision making. *Cognitive Science*, 46(5):e13143. Publisher: Wiley Online Library.

Rash, C. J., Weinstock, J., and Van Patten, R. (2016). A review of gambling disorder and substance use disorders. *Substance abuse and rehabilitation*, 7:3. Publisher: Dove Press.

Ratcliff, R. and Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, 83(3):190. Publisher: American Psychological Association.

Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2):333. Publisher: American Psychological Association.

Recanatesi, S., Pereira-Obilinovic, U., Murakami, M., Mainen, Z., and Mazzucato, L. (2022). Metastable attractors explain the variable timing of stable behavioral action sequences. *Neuron*, 110(1):139–153. Publisher: Elsevier.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3):147. Publisher: Nature Publishing Group.

Reutskaja, E., Nagel, R., Camerer, C. F., and Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, 101(2):900–926.

Rich, E. L. and Wallis, J. D. (2016). Decoding subjective decisions from orbitofrontal cortex. *Nature Neuroscience*, 19(7):973–980. Publisher: Nature Publishing Group.

Roberts, J. H. and Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4):429–440. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, 108(2):370. Publisher: American Psychological Association.

Rubin, J., Shamir, O., and Tishby, N. (2012). Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., and Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6):1054–1069. Publisher: Elsevier.

Russell, S. and Wefald, E. (1991). Principles of metareasoning. *Artificial intelligence*, 49(1-3):361–395. Publisher: Elsevier.

Russell, S. J. and Subramanian, D. (1994). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67. Publisher: Elsevier.

Santos, B., Hortaçsu, A., and Wildenbeest, M. (2012). Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior. *American Economic Review*, 102:2955–2980.

Scheibehenne, B., Greifeneder, R., and Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, 37(3):409–425. Publisher: The University of Chicago Press.

Schmidhuber, J. (1991a). Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463.

Schmidhuber, J. (1991b). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., and others (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609. Publisher: Nature Publishing Group.

Schulman, J., Chen, X., and Abbeel, P. (2017). Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*.

Schwartz, M., Sadler, P., Sonnert, G., and Tai, R. (2009). Depth versus breadth: How content coverage in high school science courses relates to later success in college science coursework. *Sci. Ed*, 93:798–826.

Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. (2021). State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pages 9443–9454. PMLR.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., and De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, 9:e60705. Publisher: eLife Sciences Publications Limited.

Shadlen, M. and Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, 90:927–939.

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. (2020). Dynamics-Aware Unsupervised Skill Discovery. *International Conference on Learning Representations*. MAG ID: 2995736683 S2ID: ae3b2768b0a3c73410bce0d2ae03feaf01f6f864.

Shenhav, A., Barrett, L., and Bar, M. (2013). Affective value and associative processing share a cortical substrate. *Cognitive, Affective, & Behavioral Neuroscience*, 13:46–59 10 3758 13415–012–0128–4.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., and Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40:99–124. Publisher: Annual Reviews.

Shocker, A. D., Ben-Akiva, M., Boccara, B., and Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, 2(3):181–197. Publisher: Springer.

Siegel, M., Buschman, T. J., and Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, 348(6241):1352–1355. Publisher: American Association for the Advancement of Science.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1):99–118. MAG ID: 2148962857.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138. MAG ID: 1989388297.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, 1(1):161–176. Publisher: North-Holland.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Sims, C. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50:665–690 10 1016 0304–3932 03 00029–1.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, 152:181–198. Publisher: Elsevier.

Singh, S., Lewis, R. L., and Barto, A. G. (2009). Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, pages 2601–2606. Cognitive Science Society.

Sleezer, B. J., Castagno, M. D., and Hayden, B. Y. (2016). Rule encoding in orbitofrontal cortex and striatum guides selection. *Journal of Neuroscience*, 36(44):11223–11237. Publisher: Soc Neuroscience.

Sobel, M. and Huyett, M. (1957). Selecting the Best One of Several Binomial Populations. *Bell System Technical Journal*, 36:537–576 10 1002 1538–7305 1957 02411.

Sterling, P. and Laughlin, S. (2017). *Principles of Neural Design*. MIT Press. Google-Books-ID: HA78DwAAQBAJ.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225. Publisher: The University of Chicago Press.

Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148. Publisher: Springer.

Sutton, R. S., Barto, A. G., and others (1998). Introduction to reinforcement learning. Publisher: MIT press Cambridge.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211. MAG ID: 2109910161 S2ID: 0e7638dc16a5e5e9e46c91272bfb9c3dd242ef6d.

Tajima, S., Drugowitsch, J., Patel, N., and Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, 22(9):1503–1511. Publisher: Nature Publishing Group.

Tajima, S., Drugowitsch, J., and Pouget, A. (2016). Optimal policy for value-based decision-making. *Nat Commun*, 7:12400.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Xi Chen, O., Duan, Y., Schulman, J., DeTurck, F., and Abbeel, P. (2017). # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30.

Thomas, A. W., Molter, F., and Krajbich, I. (2021). Uncovering the computational mechanisms underlying many-alternative choice. *Elife*, 10:e57012. Publisher: eLife Sciences Publications Limited.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, 25(3):219–225. Publisher: Wiley Online Library.

Timmermans, D. (1993). The impact of task complexity on information use in multi-attribute decision making. *Journal of Behavioral Decision Making*, 6:95–111.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, 20(4):410–433. Publisher: Wiley Online Library.

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

Todorov, E. (2006). Linearly-solvable Markov decision problems. *Advances in neural information processing systems*, 19.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483. Publisher: National Acad Sciences.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1295–1306. Publisher: The Royal Society.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, 76(3):282. Publisher: American Psychological Association.

Turner, S., Bettis, R., and Burton, R. (2002). Exploring Depth Versus Breadth in Knowledge Management Strategies. *Computational & Mathematical Organization Theory*, 8:49–73.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79:281–299.

Usher, M. and McClelland, J. (2004). Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice. *Psychological Review*, 111:757–769 10 1037 0033–295 111 3 757.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, 108(3):550. Publisher: American Psychological Association.

Vickery, T. J., Chun, M. M., and Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, 72(1):166–177. Publisher: Elsevier.

Vidal, A., Soto-Faraco, S., and Moreno-Bote, R. (2022). Balance between breadth and depth in human many-alternative decisions. *eLife*, 11:e76985. Publisher: eLife Sciences Publications, Ltd.

Volpi, N. C. and Polani, D. (2020). Goal-directed Empowerment: combining Intrinsic Motivation and Task-oriented Behaviour. *IEEE Transactions on Cognitive and Developmental Systems*. Publisher: IEEE.

von Neumann, J. and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press. MAG ID: 2144846366.

Vul, E., Goodman, N. D., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637. MAG ID: 2157201325.

Wang, M. Z. and Hayden, B. Y. (2021). Latent learning, cognitive maps, and curiosity. *Current Opinion in Behavioral Sciences*, 38:1–7. Publisher: Elsevier.

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234. Publisher: Elsevier.

Wilson, R., Geana, A., White, J., Ludvig, E., and Cohen, J. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143:2074–2081.

Wilson, R. C., Bonawitz, E., Costa, V. D., and Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences*, 38:49–56. Publisher: Elsevier.

Wissner-Gross, A. D. and Freer, C. E. (2013). Causal entropic forces. *Physical review letters*, 110(16):168702. Publisher: APS.

Yantis, S. and Johnston, J. C. (1990). On the locus of visual selection: Evidence from focused attention tasks. *Journal of Experimental Psychology: Human perception and performance*, 16(1):135. Publisher: American Psychological Association.

Yoo, S. B. M. and Hayden, B. Y. (2018). Economic choice as an untangling of options into actions. *Neuron*, 99(3):434–447. Publisher: Elsevier.

Zhang, C., Cai, Y., Huang, L., and Li, J. (2021a). Exploration by maximizing Rényi entropy for reward-free RL framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10859–10867. Issue: 12.

Zhang, T., Rosenberg, M., Perona, P., and Meister, M. (2021b). Endotaxis: A Universal Algorithm for Mapping, Goal-Learning, and Navigation. *bioRxiv*. Publisher: Cold Spring Harbor Laboratory.

Ziebart, B. D. (2010). *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University.

Zylberberg, A., Dehaene, S., Roelfsema, P. R., and Sigman, M. (2011). The human Turing machine: a neural framework for mental programs. *Trends in Cognitive Sciences*, 15(7):293–300. Publisher: Elsevier.

Ágh, T., Kovács, G., Supina, D., Pawaskar, M., Herman, B. K., Vokó, Z., and Sheehan, D. V. (2016). A systematic review of the health-related quality of life and economic burdens of anorexia nervosa, bulimia nervosa, and binge eating disorder. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 21(3):353–364. Publisher: Springer.