Thinking and Guessing: Bayesian and Empirical Models of How Humans Search

Marta Kryven (mkryven@uwaterloo.ca)
Department of Computer Science,
University of Waterloo

Tomer Ullman (tomeru@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

William Cowan (wbcowan@uwaterloo.ca)
Department of Computer Science
University of Waterloo

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Searching natural environments, as for example, when foraging or looking for a landmark in a city, combine reasoning under uncertainty, planning and visual search. Existing paradigms for studying search in humans focus on its isolated aspects, such as step-by-step information sampling or visual search, without examining advance planning. We propose and evaluate a Bayesian model of how people search in a naturalistic maze-solving task. The model encodes environment exploration as a sequential process of acquiring information modelled by a Partially Observable Markov Decision Process (POMDP), which maximises the information gained. We show that the search policy averaged across participants is optimal. Individual solutions, however, are highly variable and can be explained by two heuristics: *thinking* and *guessing*. Self-report and inference using a Gaussian Mixture Model over inverse POMDP consistently assign most subjects to one style or the another. By analysing individual participants' decision times during the task we show that individuals often solve partial POMDPs and plan their search a limited number of steps in advance.

Keywords: spatial search; exploration; Inverse Bayesian Inference; Partially Observable Markov Decision Process; decision-making;

Introduction

Exploring has always been part of humanity's larger story, from scouring the savannah for food to navigating the scattered Polynesian islands. However, goal-directed exploration is also a common daily activity. Imagine an exhibition visitor looking for a particular painting. Her map shows the size and location of the gallery rooms, but not where specific paintings can be located (as, for example, on Figure 1, top). Which route should she take to find the painting? Or consider a tourist in a busy market

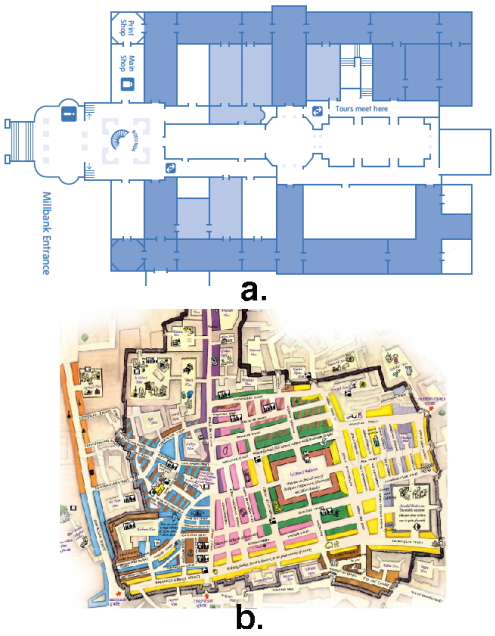

Figure 1: Examples of searching in a natural environment. (a.) Tate Britain gallery. (b.) Istanbul bazaar.

who wants to find the best presents for his family, at a reasonable price. Given limited time, how should he search the market (Figure 1, bottom)?

Shopping, foraging for food, or discovering a new ocean are all examples of everyday search, where the agent has partial knowledge of its environment and acts to gain information extending its knowledge, while seeking reward and minimizing cost. The desire to search can come about from the belief that unexplored spaces may contain a higher reward than is currently available. However, it appears that exploration is also driven by an intrin-

sic reward for novelty (Dayan & Sejnowski, 1996). There is also evidence that searching in an abstract domain is similar to searching in a physical space, and that search behaviors are are consistent between individuals. For example, priming strategies of spatial search affects how humans search for words in memory (Hills et al., 2008).

Intuitively, a rational agent searches by combining planning, learning, and reasoning under uncertainty to maximise its knowledge about the environment. Consider the gallery visitor looking for a specific painting. The visitor has a map of the exhibition space, a view of a few gallery rooms and a memory of viewing others. In this state, she must inspect each room and hallway in turn to find out if the painting is there. Since larger spaces generally contain more paintings, the visitor may prefer larger rooms. However, chancing upon a small room at the side, she will likely take the small cost to visit it, to avoid backtracking later (Figure 2).

A majority of human observers are quite good at interpreting searching behavior by other agents and judge actions generated from a rational planning strategy as more intelligent than actions generated from other policies. However, a large minority consistently attribute intelligence to the outcome rather than strategy ('if she found the painting, she must've done something right, regardless of how she got there') (Kryven et al., 2016).

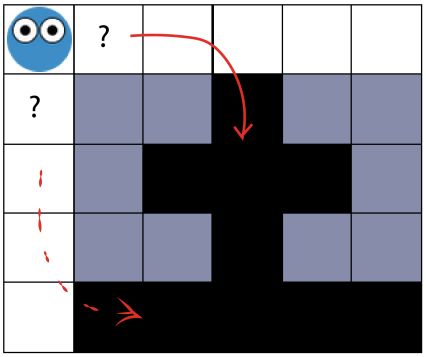

Figure 2: Different possible paths: The solid arrow shows the optimal exploration path when looking for a target, maximising IG over time, while the dashed arrow shows a suboptimal path.

Do people plan ahead rationally when they search? Many formal approaches consider search as a step-by-step information sampling process, with little attention to planning in advance beyond a single step (Markant et al., 2016; Markant & Gureckis, 2014; Najemnik & Geisler, 2005; Nelson et al., 2010). In this study we draw on work in robotics, and formalize everyday search as a Partially Observable Markov Decision Process (POMDP), similar to a model of how people think about other's planning (Kryven et al., 2016). We examine our framework using an empirical escape-the-maze task, in which human participants searched for hidden goal in different maze layouts.

Our results indicate that people maximise the information gained over several steps ahead, with individual differences in how deliberately they search. In the following sections we discuss related work, followed by a model-based account of optimal search and its experimental validation. Empirical analysis shows how the model can be extended to reflect search in resource-limited domains.

Formal Models of Search

Formally, search can be captured in different ways. Searching hypothesis spaces in active learning (Markant et al., 2016) and psychophysical Bayesian Ideal Observer models (Najemnik & Geisler, 2005) reward an agent's reduction in uncertainty by a common principle of Information Gain (IG), maximizing the information gained on the next time step to best adjudicate between available hypotheses. Novelty preference models, common in robotics, directly give the agent small rewards for taking novel actions (Brafman & Tennenholtz, 2002; Bellemare et al., 2016).

The model that comes closest to searching a natural environment is Bayesian Ideal Observer (Najemnik & Geisler, 2005), which shows that humans search for a visual target embedded in noise by maximising the IG of each eye-movement, but without integrating partial samples across successive saccades. Even so, human and primate searchers achieve near-optimal performance thanks to a phenomenon of *inhibition of return*, a tendency to avoid recently fixated display locations. Thus, the decisions about where to sample can be explained by a self-avoiding one step ideal observer who re-

tains a memory for fixated locations.

Studies of searching hypotheses spaces in active learning assume reasoning step-by-step, asking questions like: 'How many alternative hypothesis do people evaluate at a time?' or 'How do people incorporate unreliable evidence?' (Markant et al., 2016). Normative IG models consider the expected reduction in uncertainty about the true hypothesis across all possible outcomes of a query (Nelson et al., 2010; Markant & Gureckis, 2014). However, in complex tasks humans may act in accordance with a simpler model, adjudicating between just two alternatives at a time (Markant et al., 2016) and using information from multiple hypotheses gradually, depending on available mental resources.

To study everyday search across multiple time steps, we use a maze-world experimental paradigm in which information is disclosed progressively, in effect extending the task described in (Kryven et al., 2016).

Computational Framework

We model search as probabilistic planning in a POMDP. This framework assumes that the agent acts sequentially to maximise the reward and minimise the cost of each action, given its beliefs about the world. After each action the agent updates its beliefs based on observations caused by the previously chosen action. The family of POMDP models can describe any behaviour encoded as a discretised sequential process and share a common principle of Bayesian belief updating. In theory the model can admit a variety of cost functions, rewards, reward discount rates and observation models, depending on the modelled problem.

Consider an agent looking for an exit in a 2-D maze (as shown in Figure 2). The agent has a partial knowledge of the maze: it knows the size of the rooms, and how accessible they are. It knows there is an exit in the maze, but does not know where it is. The agent has a 180 degree view, such that the cells not yet seen by the agent remain dark, each equally likely to hide the exit. To search efficiently, the agent decides which rooms to sample first to best narrow down probabilities over locations of the exit.

For concreteness we assume reward as a function of what is known about the world with an added exploration bonus. The bonus ensures a small probability of taking any action, even if unrewarding, adding flexibility to possible behaviours.

Formally, we based the agent's planning on (Kryven et al., 2016). The world is described by discrete time, $0 \leq t \leq T$ and a grid of cells $W = \{w(i,j)\}, \quad w(i,j) \in \{wall, empty, goal\}$. The agent's beliefs are a set of probabilities $X_t = \{P(\mathbf{x_s})_t\}$, over world states $\{\mathbf{x_s}\}$ and X_0 encodes the set of the agent's initial beliefs. On each time step the agent at a location L_t receives new observation probabilities $O_t = P(W|X_t, L_t)$, updates its beliefs using standard Bayesian updating: $X_{t+1} = P(W|A_t, X_t) \propto P(O_t|A_t, X_t)P(A_t|X_t)P(W)$, and chooses among actions $A(L_t) = \{a_i\} \in \{N, S, W, E\}$ with a likelihood proportional to the action's reward $R : X_t \times A_t \mapsto \mathbb{R}$.

$$R(Q_t(a_i)) = \frac{exp(Q_t(a_i)/\tau)}{\sum_j exp(Q_t(a_j)/\tau)} + \varepsilon, \qquad (1)$$

where $Q(a, L_t, X_t)$ reflects the value of information the agent expects to learn about the environment. τ is a $softmax$ parameter controlling decision noise and ε controls the exploration bonus. The agent's reward and observation models are based based on (Kryven et al., 2016). This model describes an optimal search, which provides a useful benchmark to compare with people's actual search behavior, to which we turn next.

Experiments

Participants 120 participants were recruited via Amazon Mechanical Turk, 46 female and 74 male, mean age 33, SD=10.13. 60 participants did the experiment in the Bonus condition, in which the top 20% of participants received a bonus for finishing with the overall lowest step cost. The remaining 60 in the No Bonus condition received no bonus.

10 participants were excluded for failing to answer questions and five more answers were generated by the same person with multiple MTurk accounts. The procedure received ethics clearance from a University of Waterloo Research Ethics Committee, and from an MIT Ethics Review Board.

Stimuli and Procedure Participants were instructed to find a hidden goal location ('exit') in a series of mazes, by controlling an animated character. The character can move one grid square at a time: N, W, S, E and has a 180 degree view limited by walls. The maze is initially dark, but is uncovered as the character moves along, so participants initially know the layout of the rooms, but not where the goal is (marked as a bright red circle once in line of sight). Participants are instructed that each of the dark squares is equally likely to hide the 'exit', and that they should find it in as few steps as possible.

After reading the instructions, participants solved three practice mazes and answered instruction-comprehension questions, followed by 12 more mazes. At the end of the experiment participants are were asked how they made their decisions. The participants' decision times and their path through the maze were recorded on each trial. The full experimental procedure is available at `http://cgl.uwaterloo.ca/~mkryven/`.

Results Each solution was labeled according to the most likely POMDP settings estimated by the inverse-planning inference: *optimal*, *softmax* or *softmax-expl*. Here *softmax* indicates a solution generated by an agent with $\tau > 0$ (we used $\tau \in [0.01, 0.1]$) and *expl* indicates $\varepsilon > 0$. Additionally, we calculated the fraction of optimal steps, moves consistent with the optimal POMDP solution starting in the same state, taken by each participant.

There was no significant difference between the mean fractions of optimal steps $t = -0.6204, p = .5$ or mean decision times (DT) $t = -0.0644, p = .9$ of participants in the Bonus and the No Bonus conditions, so we collapsed the conditions. Compared to the No Bonus condition *exploration bonus* in the Bonus condition was reduced (Table 1), so participants who received a bonus made fever unrewarding moves and paid more attention to instructions.

14 participants (4 participants from the Bonus condition and 10 from the No Bonus condition) failed an attention check procedure, scoring a more than 5 *softmax-expl* trials. Solutions labelled as *softmax-expl* capture solutions with motor mistakes or unusual strategies, such as entering empty rooms, following walls or clicking at random. Thus, we reasoned that participants who generated many of such solutions were inattentive.

Table 1: Model-based inference over individual trials.

Condition	Optimal	Softmax	Softmax+Expl
Bonus	45%	48%	7%
No Bonus	45%	39%	16%

Next, we calculated the general empirical policy, a table of probabilities that participants take each action (N,E,S,W) in each of the visited states. The idea of empirical policy comes from machine learning, where robotic navigation is often solved by a policy for action selection in each of the possible system states. Remarkably, in all 12 mazes the path from the starting square following the most likely actions until the goal was identical to the optimal solution (e.g. the solid line solution shown on Figure 2). Thus, solutions chosen by the majority (if participant could vote on each step) were always optimal. Individual solutions, however, were often sub-optimal (e.g. the dashed line on figure 2), so that individuals took on average 84% optimal steps ($SD = 5.9\%$).

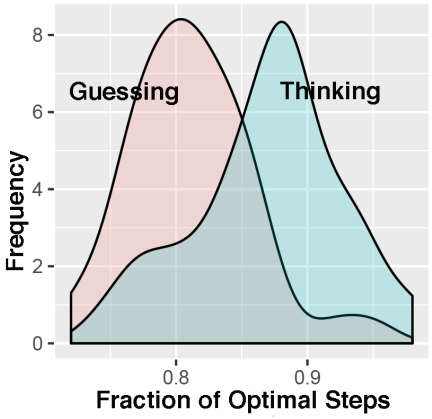

Figure 3: Participants self-describing as **thinking** were more optimal than those self-describing as **guessing**

What causes sub-optimal solutions? Since observers evaluate the intelligence of others' planning either by outcome or by strategy (Kryven et al., 2016), we hypothesised there may be a similar division in how people plan.

Using a semi-automated method (Kryven & Cowan, 2016) we determined two categories in par-

ticipants' responses to *'How did you make your decision?'* as **thinking** or **guessing**. Two independent raters using these categories agreed on 84 out of 91 participants, coding 45 of them as **thinking** and 39 as **guessing**. For example, a response was coded as **thinking** if it read: *'I tried to maximise the number of squares revealed per step.'* and as **guessing** if it read *'I followed my gut*. The remainder were randomly assigned to either group. Participants who self-described as **thinking** were on average more optimal ($t = 2.574, p = .01$ and Figure 3).

Independent estimates obtained by a Gaussian Mixture Model (GMM) over fractions of optimal steps identify the two peaks coinciding with **thinking** and **guessing** means. Both Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) prefer a GMM with two components ($\mu_1 = .86$ and $\mu_2 = .78$, $AIC = -250.99, BIC = -238.44$) over three ($\mu_1 = .87, \mu_2 = .79$ and $\mu_3 = .94$, $AIC = -246.78, BIC = -226.69$).

Possibly, people guess to minimise cognitive effort. However, since guessers may sometimes do better than optimal, some may gamble on suboptimal moves intentionally, which can be modelled by $\tau > 0$ or by a biased prior belief X_0.

Model-Free validation To validate the model and differentiate between causes of guessing we analysed participant's DT as a measure of cognitive effort. DT were pre-processed to remove data-points longer than 10s and outliers more than 3 standard deviations away from the mean, discarding 1.4% of measurements. We normalised DT by mean-dividing to remove the effect of participant, obtaining a DT distribution with a mean at 690.8ms and a median at 530ms.

Informally inspecting DT at different maze locations revealed that participants are faster in some locations than in others. Therefore, we coded each square along each trajectory by a pre-processing script as shown on figure 4. DT distribution densities in each type of squares are shown on figure 5. B squares are not shown, since the B-distribution is identical to the N-distribution delayed by 20.67ms ($t = -5.0216, p <= .0001$).

Fully solving a POMDP requires computing a solution at the start of the trial and following it until the end. So if participants solve a POMDP fully then DT should be large at the starting square and increase slightly at observation points (O and G on figure 4) by the time it takes to perform easy visual search. If, however, participants are solving a POMDP partly, DT at ordinary observations points (O squares) should be longer compared to goal observations points (G squares).

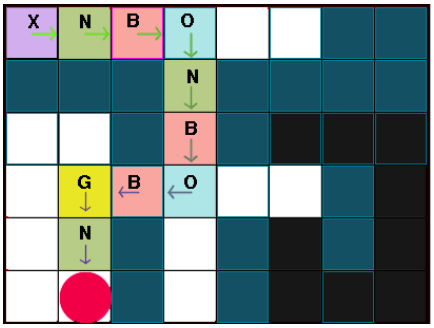

Figure 4: Maze locations were coded as: X-the starting step, O - observation points, N - neutral squares, B - before an observation point and G - observations from where the goal was observed.

DT distributions contradict the hypothesis that participants solve POMDP fully (figure 5). Although, DT in the starting square are the longest, DT in G and O squares are significantly different according to a Kolmogorov-Smirnov test ($D = 0.18, p < .0001$) and a lot longer than expected from an easy visual search, in support of a partial POMDP solution.

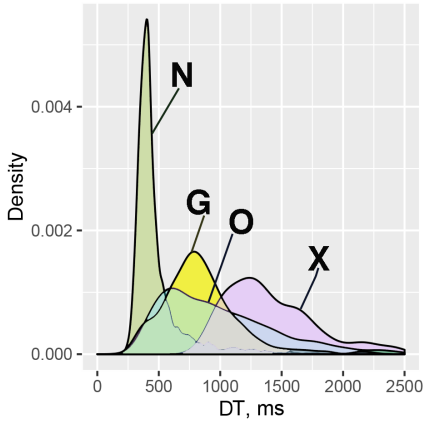

Figure 5: Decision time distribution densities across different maze locations.

Discussion

We proposed and evaluated a Bayesian model of how humans search in a naturalistic maze-solving task, which requires combining visual search, planning, and reasoning under uncertainty. The model encodes search behaviour as a sequential process of acquiring information conveniently modelled by a POMDP, which maximises the information gained.

The model predicts solutions averaged across participants, and offers functional causes for variance in individual solutions: *softmax* rewards, *exploration bonus*, and biased prior beliefs. When classifying participants' self-reports of their strategy as **thinking** or **guessing**, more deliberate planning was captured as more optimal by the model.

Model-free analysis, however, reveals a distribution of effort inconsistent with fully solving a POMDP. At least some suboptimal solutions are caused by participants planning a limited the number of steps ahead, minimising cost and effort in addition to maximising information. Some suboptimal strategies may also be intentional: Participants may gamble on an unlikely outcome, or have a complex model assuming that the goal is hidden by an adversary. We are currently investigating possible causes of suboptimal planning, as well as how many steps ahead people plan.

Our model advances the study of how people search, and offers a formal framework in which it can be explored. While our study focused on a simple spatial task, everyday search and exploration is much more than a searching in a physical space. The imperative that drives explorers to search the depths of the ocean also motivates an artist to search abstract spaces of images, or a child to explore possible explanations. Abstracting search formally takes a step toward conceptualising the common cognitive process that animates human curiosity and exploration across domains.

Acknowledgments

We thank members of MIT CoCoSci for valuable discussion. This work was supported by Center for Minds, Brains and Machines under NSF STC award CCF-1231216.

References

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Proceedings of neural information processing 2016.*

Brafman, R. I., & Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*(Oct), 213–231.

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*(1), 5–22.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, *19*(8), 802–808.

Kryven, M., & Cowan, W. (2016). Semi-automated classification of free-form participant comments. In *Proceedings of wiml workshop at neural information processing 2016.*

Kryven, M., Ullman, T., Cowan, W., & Tenenbaum, J. B. (2016). Outcome or strategy? a bayesian model of intelligence attribution. In *Proceedings of the thirty-eighths annual conference of the cognitive science society.*

Markant, D. B., & Gureckis, T. M. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 958–963).

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, *40*(1), 100–120.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters information acquisition optimizes probability gain. *Psychological science*, *21*(7), 960–969.