Planning to plan: a Bayesian model for optimizing the depth of decision tree search

Ionatan Kuperwajs (ikuperwajs@nyu.edu)
Center for Neural Science, New York University
New York, NY, United States

Wei Ji Ma (weijima@nyu.edu)
Center for Neural Science and Department of Psychology, New York University
New York, NY, United States

Abstract

Planning, the process of evaluating the future consequences of actions, is typically formalized as search over a decision tree. This procedure increases expected rewards but is computationally expensive. Past attempts to understand how people mitigate the costs of planning have been guided by heuristics or the accumulation of prior experience, both of which are intractable in novel, high-complexity tasks. In this work, we propose a normative framework for optimizing the depth of tree search. Specifically, we model a metacognitive process via Bayesian inference to compute optimal planning depth. We show that our model makes sensible predictions over a range of parameters without relying on retrospection and that integrating past experiences into our model produces results that are consistent with the transition from goal-directed to habitual behavior over time and the uncertainty associated with prospective and retrospective estimates. Finally, we derive an online variant of our model that replicates these results.

Keywords: sequential decision-making; planning; Bayesian inference

Introduction

From spatial navigation to organizational strategy to playing Go, planning is a hallmark of human intelligence. Planning involves the mental simulation of future actions and their consequences in order to make a decision. However, evaluating every possible course of action in complex environments is simply intractable. For example, if an agent has to make a sequence of N decisions with K options at each step, then the total number of sequences is K^N.

Planning problems have typically been formalized as search over a decision tree in both cognitive science (Daw, Niv, & Dayan, 2005; Huys et al., 2015; van Opheusden et al., 2021) and artificial intelligence (Shannon, 1950; Silver et al., 2016). In such a scheme, the agent builds a tree of possible future trajectories where every decision that the agent must make is represented by a branching point. The agent then gains information by traversing the decision tree, which is used to estimate the long-term expected reward of each currently available action.

Tree search algorithms generally lead to better decisions that may have been overlooked without planning, but can be costly to run. Even with a small cost per unit of time or planning iteration, real-world tasks involve too many possible actions to extensively evaluate each one considering the breadth and depth of the trees an agent would need to construct. Therefore, a growing body of literature has focused on solutions for approximating the values of choices without fully expanding a search tree (Snider, Lee, Poizner, & Gepshtein, 2015) or efficiently allocating limited computational resources during planning (Callaway et al., 2018). Other mechanisms in planning models that achieve similar goals involve pruning initially unpromising courses of action (Huys et al., 2012), relying on the uncertainty or accuracy of forward search and model-free reinforcement learning methods in tandem (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Kool, Gershman, & Cushman, 2017; Hamrick et al., 2019), or leveraging simulated experience to further expedite the transition from goal-directed to habitual behavior (Dasgupta, Schulz, Goodman, & Gershman, 2018). Meanwhile, simpler choice models of human planning often do not explore the tradeoffs between the costs of planning and decision quality (Solway & Botvinick, 2015; Tajima, Drugowitsch, Patel, & Pouget, 2019).

As a result, previous models of human planning have been predominantly guided by researcher-specified heuristics. A notable exception to this computes the value of information gained by planning in a principled manner (Sezener, Dezfouli, & Keramati, 2019), but relies on a habitual system to estimate values at the frontier of the search tree. Such a method may not scale well to complex tasks where an agent almost exclusively encounters unique states, thus hindering the development of informative habits.

Here, we propose an alternative: a normative Bayesian model for optimizing the depth of decision tree search. Our framework is inspired by perceptual cue combination models, and operates at a metacognitive level of abstraction by asking how far in advance an agent should plan before any planning actually takes place. This method has the added capability to incorporate retrospective experience as it becomes available to provide better depth estimates, but does not rely on model-free state learning in order to make sensible predictions. We make a number of simplifying assumptions that can be replaced by more sophisticated planning methods, but we note that the purpose of a metacognitive algorithm is to approximate expected reward while simultaneously reducing costs. Our simulation results show that this framework derives intuitive principles about the depth to which planning is beneficial as a function of the cost per measurement, the total number of actions the agent must evaluate, the amount of accumulated retrospective experience, and the uncertainties associated with both prospective and retrospective samples.

We conclude by deriving an online version of the model and discussing its cognitive plausibility when applied to simple and complex planning tasks.

Model

From a metacognitive perspective, an agent must think about how far into the future it may be beneficial to plan in a given state. To address this, assume the agent is in state s with actions $\{a_1, ..., a_n\}$ available and has the option of executing a tree search policy π. The key insight that our model is based on is that each state-action pair has a theoretical long-running expected reward under π, $Q(a)$. We take a Bayesian inference view where this value is unknown to the agent, and the agent tries to build a probability distribution over each $Q(a)$ while balancing the costs of search. This distribution is primarily computed by considering how a one-step, myopic evaluation will change under a certain depth of search, and is combined with any prior experience that the agent has already accumulated for each state-action pair. We reiterate that this process occurs prior to any forward search taking place, and that this paper introduces the general framework for approximating the effects of planning under the agent's particular tree search policy.

Generative model

To simulate the measurements that the agent has available to them prior to planning (Figure 1A), we assume that the true value Q for a given state-action pair is normally distributed:

$$Q \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right). \tag{1}$$

We omit the dependence of the parameters on the action a from the notation until the last step, where we will compute the value of the state across all actions. A retrospective experience with a is modeled as a noisy measurement, $q_{\text{retro},i}$, drawn from a normal distribution centered at the true Q:

$$q_{\text{retro},i} \sim \mathcal{N}\left(Q, \sigma_{\text{retro}}^2\right). \tag{2}$$

The retrospective measurements form a vector $\mathbf{q}_{\text{retro}} \equiv (q_{\text{retro},1}, \ldots, q_{\text{retro},n})$, where n is the number of past experiences with action a in state s. Similarly, the agent can perform a one-step look-ahead to obtain another noisy measurement, q_1, of Q:

$$q_1 \sim \mathcal{N}\left(Q, \sigma^2\right). \tag{3}$$

The core part of our framework is a statistical model maintained by the agent of the effects of prospective tree search without actually performing the search. We assume that each iteration of the tree search algorithm works on a branch that starts with action a and produces a new, independent measurement of Q, q_t. Therefore, after T iterations of tree search, the agent has another vector of measurements $\mathbf{q} \equiv (q_1, \ldots, q_T)$.

Inference

The overarching goal of this framework is to solve for the optimal number of iterations to plan for, T^*. We take a normative approach, meaning that we assume the agent makes this decision by maximizing expected reward given costs. T is optimized independently of any action, which includes the possibility of choosing $T = 1$, or no planning at all. One way to conceptualize this is that, before planning, the agent approximates a breadth-first search algorithm by evaluating the future expected reward at each action to the same depth.

The general inference scheme is outlined in Figure 1B and works as follows: (1) the agent considers different futures for each action after T planning iterations, (2) this future distribution is integrated with any retrospective information to form a posterior distribution, and (3) the agent marginalizes over all possible futures by combining across actions and computing the maximum over the distribution of the posterior's expected value. The final output of this inference procedure is the mean of the max distribution, which we call the value of planning for T iterations.

Formally, the posterior is the normalized product of the prior, the retrospective likelihood, and the prospective likelihood, all of which we assume to be independent:

$$p(Q|\mathbf{q}_{\text{retro}}, \mathbf{q}) \propto p(Q)p(\mathbf{q}_{\text{retro}}|Q)p(\mathbf{q}|Q). \tag{4}$$

Each of the likelihoods is over Q based on the retrospective or prospective measurements available to the agent:

$$p(\mathbf{q}_{\text{retro}}|Q) = \mathcal{N}\left(Q; \bar{q}_{\text{retro}}, \frac{\sigma_{\text{retro}}^2}{n}\right) \tag{5}$$

$$p(\mathbf{q}|Q) = \mathcal{N}\left(Q; \bar{q}, \frac{\sigma^2}{T}\right), \tag{6}$$

where $\bar{q}_{\text{retro}} \equiv \sum_{i=1}^{n} q_{\text{retro},i}$ and $\bar{q} \equiv \sum_{t=1}^{T} q_t$. This allows us to rewrite the posterior as the normal distribution $\mathcal{N}\left(Q; \mu_T, \sigma_T^2\right)$, where we define the mean and variance as

$$\mu_T = \frac{J_0\mu_0 + J_{\text{retro}} n\bar{q}_{\text{retro}} + JT\bar{q}}{J_T} \tag{7}$$

$$\sigma_T^2 = \frac{1}{J_T}, \tag{8}$$

along with the precision quantities $J_{\text{retro}} \equiv \frac{1}{\sigma_{\text{retro}}^2}$, $J \equiv \frac{1}{\sigma^2}$, $J_0 \equiv \frac{1}{\sigma_0^2}$, and $J_T \equiv J_0 + J_{\text{retro}} n + JT$. We then write $\bar{q}$ to indicate that the myopic values, q_1, are known while the remaining values are defined as $\bar{q}_{>1} = \frac{1}{T-1}\sum_{t=2}^{T} q_t$. Now, we calculate the distribution over $\bar{q}_{>1}$ given q_1 and $\mathbf{q}_{\text{retro}}$ by marginalizing

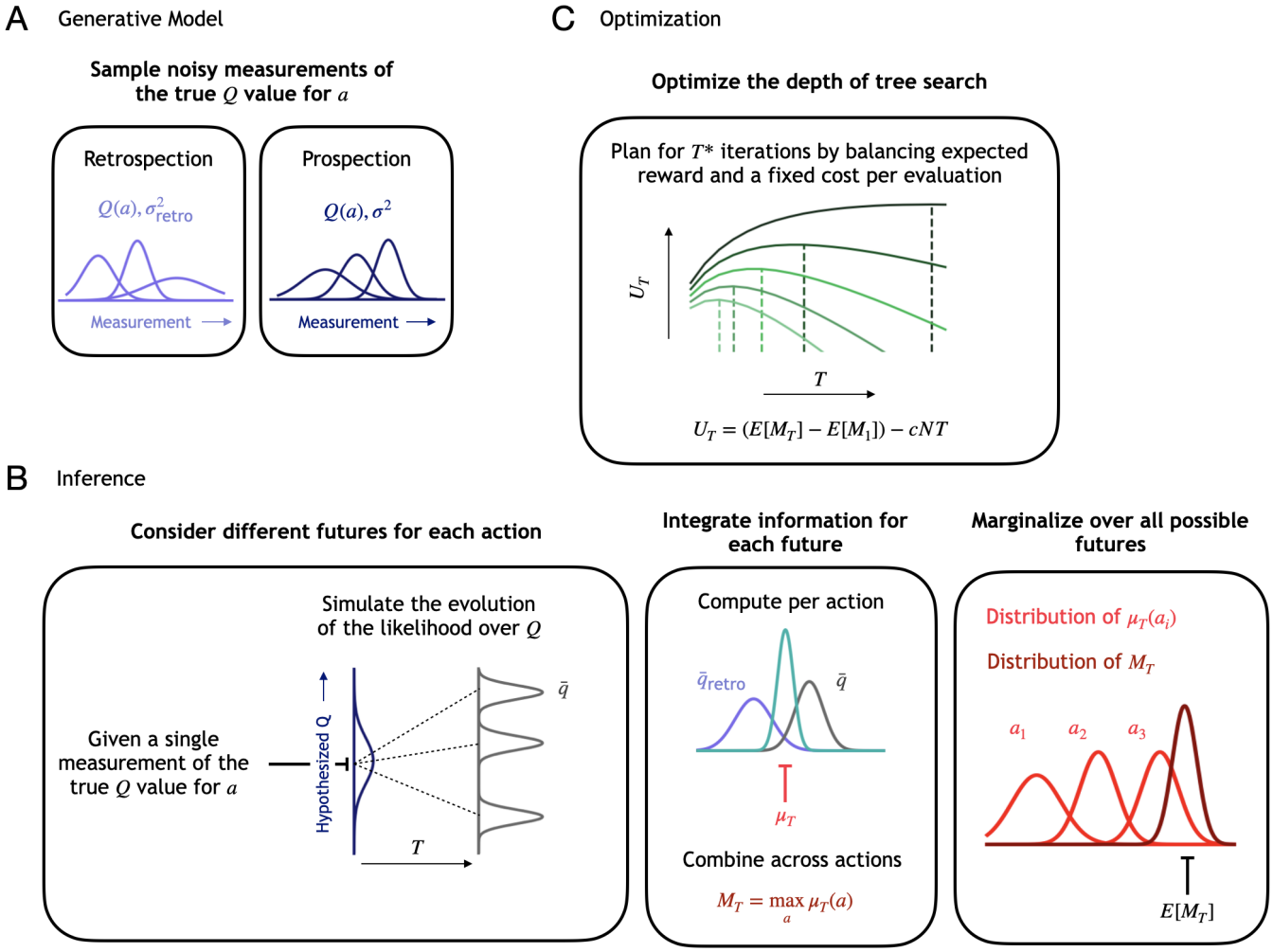

Figure 1: Planning to plan as Bayesian inference. **(A)** The generative model, where the agent receives noisy measurements of the underlying Q value for a given state-action pair. These measurements can come from retrospective experience or a myopic evaluation one step into the future, with a specific mean and variance. **(B)** The inference procedure, where the agent computes the value of the state for T planning iterations. Per action, the agent simulates the evolution of the myopic likelihood (blue distribution) T steps ahead, resulting in a sum over T prospective samples ($\bar{q}$). The gray distributions represent different potential future likelihoods. $\bar{q}$ is combined with any retrospective information (the sum over retrospective samples, $\bar{q}_{\text{retro}}$) into a posterior. The agent then combines across actions by marginalizing over all possible futures, resulting in a distribution for each posterior mean μ_T (light red distributions) and a max distribution for the state (dark red distribution). The expected value of the max distribution, $\mathbb{E}[M_T]$, is the value of planning for T. **(C)** The optimization step, where the agent repeats the inference procedure for all T. Subtracting a fixed cost per evaluation c to find the value of planning results in an optimal planning depth T^*. Note that this cost must be multiplied by the depth of search (T) and the number of actions that the agent considers (N).

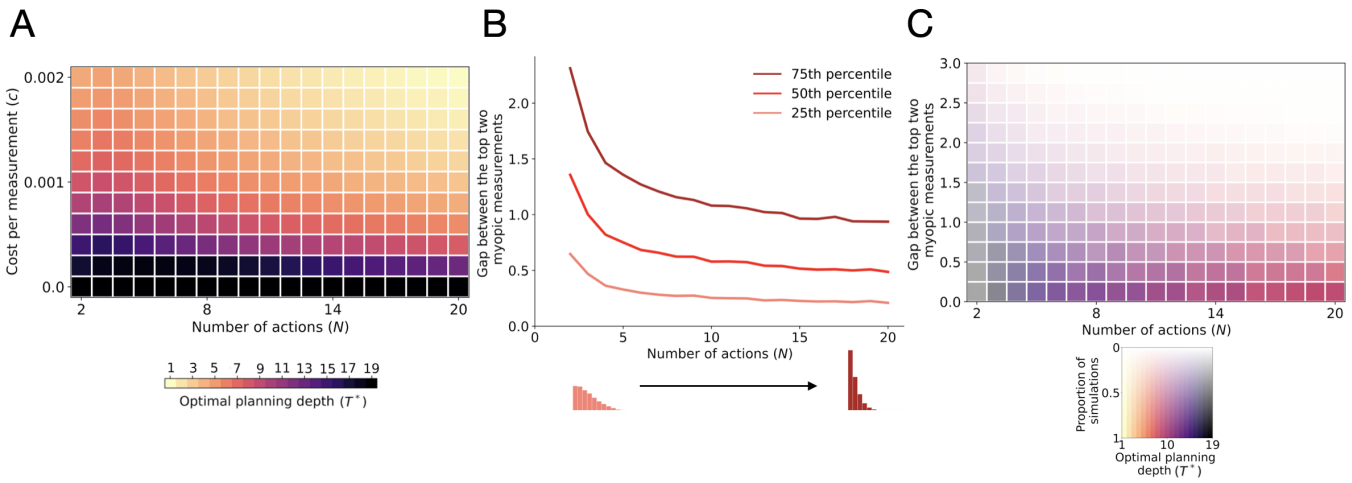

Figure 2: The model makes reasonable predictions about planning depth based solely on myopic estimates. **(A)** Optimal planning depth (T^*) as a function of the cost per measurement (c) and the number of actions available to the agent (N). **(B)** 75th, 50th, and 25th percentiles for the gap between the top two myopic measurements as a function of number of actions for a single cost ($c = 0.001$). **(C)** Optimal planning depth as a function of the gap between the top two myopic measurements and number of actions, for the same cost as in (B). The color code is two-dimensional: the hue represents optimal planning depth and the saturation proportion of total simulations for the combination of top gap size and N.

over the current possible values of Q:

$$p(\bar{q}_{>1}|q_1, \mathbf{q}_{\text{retro}}) = \int p(\bar{q}_{>1}|Q)p(Q|q_1, \mathbf{q}_{\text{retro}})dQ \quad (9)$$

$$= \mathcal{N}\left(\bar{q}_{>1}; \mu_1, \frac{\sigma^2}{T-1} + \sigma_1^2\right). \quad (10)$$

The main technical result here is computing the distribution over future prospective measurements T time steps into the future given the first myopic measurement. On average, the mean of the resultant distribution stays at the mean of the first measurement while the variance becomes narrower. The variance of this distribution at T is

$$\text{Var}[\mu_T] = \frac{\sigma_1^2}{1 + \frac{\sigma^2}{(T-1)\sigma_1^2}}. \quad (11)$$

Intuitively, the variance monotonically increases as a function of T because the future measurements are unknown and will cumulatively pull the posterior mean away from μ_1. Note that when $T = 1$, the variance is 0 as μ_1 is known and that when $T \to \infty$, the variance saturates as σ_1^2. This is because the future measurements will have fully pulled the posterior mean to the true value of Q.

The value of planning is the maximum of the future posterior mean of Q across all actions, $M_T = \max_a \mu_T(a)$. We can evaluate the expected value of this quantity, $\mathbb{E}[M_T]$, analytically by computing the max distribution and taking its mean. Within our framework, the mathematical reason why planning is beneficial is that the expected value of a maximum is greater than the maximum of expected values.

Optimization

After computing the benefit of planning for a range of T values, this expected reward is compared against the cost of planning (Figure 1C). We assume a fixed cost per evaluation c such that the utility of planning for T iterations is given by:

$$U_T = (\mathbb{E}[M_T] - \mathbb{E}[M_1]) - cNT, \quad (12)$$

where N is the total number of actions considered in the state. In this way, our cost function takes into account the depth and breadth of the tree being approximated by the model. The first term increases sublinearly with T, while the second one increases linearly, meaning that their difference will have an optimum. We numerically calculate this optimum, which is the best number of steps to plan ahead:

$$T^* = \underset{T}{\text{argmax}}\, U_T. \quad (13)$$

Results

The primary goal of this work is to allow an agent to preemptively characterize the conditions under which tree search is beneficial. We performed a set of simulations in order to validate that the model makes intuitive depth predictions.

Myopic model predictions

We first consider the case where the agent has no prior experience and relies purely on myopic evaluations to decide how far into the future to plan. This mimics real-world planning environments where an agent has uninformed priors over their retrospective system, such as in novel tasks or tasks in which states may not repeat often.

Our cost function is dependent on the number of evaluations that the agent must make during inference, or the product of the cost per measurement (c) and the total number of actions available to the agent in the state (N). Over this parameter space, our model predicts that deeper planning is beneficial at lower costs and with less alternatives (Figure 2A).

The reasoning behind this effect is based on the value of the myopic measurements. Suppose that the agent's objective is to decide between two actions. If the gap between the myopic values of these two actions is small, should the agent plan further ahead in hopes of determining which action is actually better? Or, should the agent avoid wasting valuable resources planning, since it will be unclear which action is best regardless? And conversely, if the gap between two myopic evaluations is large, should the agent plan more or less?

In Figure 2B, we show that as the number of actions increases, the distribution of the difference between the top two myopic measurements becomes more right-skewed. In other words, small top gap sizes are more common when there are more actions to consider, and the size of the top gap is varied at a lower number of alternatives. Figure 2C then investigates the relationship between top gap size and optimal planning depth: T^* is higher with smaller top gap sizes and with less available actions. This suggests that smaller gap sizes do increase optimal planning depth, but that this effect is diminished with more actions where cost grows and ultimately outweighs the added benefit of planning more deeply.

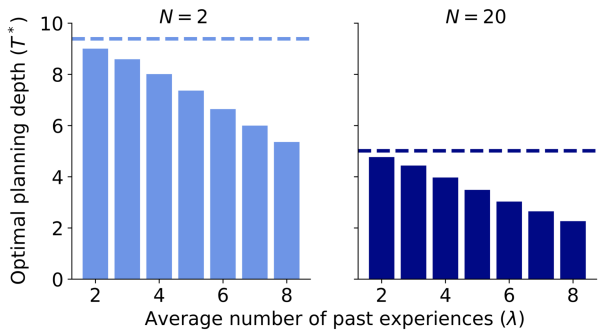

Figure 3: The effect of accumulating experience on depth of planning. Optimal planning depth for 2 actions (left) and 20 actions (right) as a function of the average number of retrospective measurements per action (modeled as a Poisson rate) for a single cost ($c = 0.001$). The dashed line in each panel indicates the optimal planning depth without retrospective experience.

Incorporating retrospective information

Next, we examine environments where the agent does have prior experience. In principle, planning depth should be modulated by the total amount of retrospective experience accumulated by the agent as well as the uncertainty of those estimates. These correlate directly to well-studied mechanisms in the planning literature: the transition from model-based to model-free control over time and uncertainty-based arbitration between prospective and retrospective systems.

We simulate total experience by using a variable Poisson rate (λ) to determine n for each action. The model predicts a shallower optimal planning depth as more experience is accumulated, and this trend holds irrespective of the number of alternatives that the agent considers (Figure 3). The rationale behind this is that environments with low amounts of retrospective information require similar planning depths compared to when the model relies only on myopic estimates. Optimal planning depth then decreases as the agent gains more experience. In these cases, the agent can spend less resources planning and instead relies more heavily on its cost-effective retrospective experiences. Additionally, we varied the amount of uncertainty for both the retrospective and myopic estimates to investigate their joint effect on planning depth (Figure 4). This is straightforward to implement, since our model directly takes the variance associated with each type of sample as part of its generative model. With a low number of alternatives, increased uncertainty with either or both sources of information led to deeper planning. With more alternatives, however, high amounts of prospective uncertainty resulted in lower optimal planning depths. Intuitively, planning more deeply is generally beneficial in gaining high-value estimates under uncertainty, but if the uncertainty attached to planning is too high then it is no longer worthwhile to obtain these costly measurements.

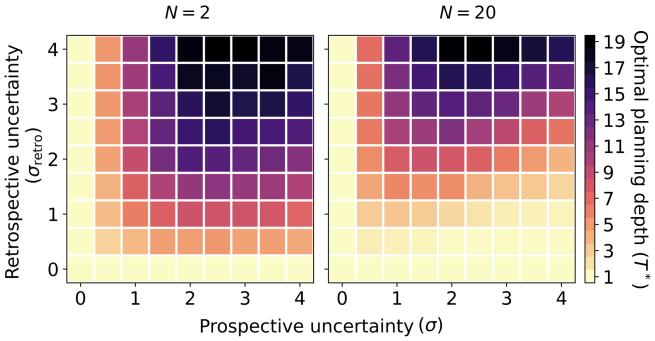

Figure 4: The effect of uncertainty on depth of planning. Optimal planning depth for 2 actions (left) and 20 actions (right) as a function of the retrospective and prospective variance in the generative model for a single cost ($c = 0.001$).

Online planning

In order to study how our framework might approximate different search procedures, we implemented a variant of the model where the agent determines online whether increasing the depth of search by one layer is worthwhile. In practice, this model could be used by the agent to iteratively learn heuristics over a set of parameters, such as a stopping rule, that are then used to inform a planning algorithm. The general logic for the model is the same: given the posterior based on retrospective and prospective information, the agent

marginalizes over possible futures. The difference is that the past measurements consist of a sequence of T measurements for each action and the future consists of a single time step, from T to $T+1$.

Mathematically, the posterior at time T is identical to Eq. (4). However, $\bar{q}_T$, which is the mean of the measurements $q_1,\ldots,q_T$, is given, rather than only q_1. These measurements may be provided by the forward search algorithm that the agent is utilizing. The new measurement q_{T+1}, which the agent receives if it plans one step further ahead, is unknown and has to be marginalized over:

$$p(q_{T+1}|\bar{q}_T,\mathbf{q}_{\text{retro}}) = \int p(q_{T+1}|Q)p(Q|\bar{q}_T,\mathbf{q}_{\text{retro}})dQ \quad (14)$$

$$= \mathcal{N}\left(q_{T+1};\mu_T,\sigma^2+\sigma_T^2\right). \quad (15)$$

Again, the mean of the distribution of the expected value of Q at $T+1$ stays at the mean of the measurement at T, while the variance at $T+1$ is

$$\text{Var}[\mu_{T+1}] = \frac{1}{J_T(1+\frac{J_T}{J})}. \quad (16)$$

Note that the variance of the online and offline implementations match when moving from $T=1$ to $T=2$. In order to combine across actions and decide whether to plan one step further, we compute the expected value of the maximum of the future posterior means for all actions and subtract a fixed cost c for that additional iteration multiplied by the number of alternatives N and the current depth of search T. This gives the utility of planning at each step as

$$U_{T+1} = (\mathbb{E}[M_{T+1}] - \mathbb{E}[M_T]) - cNT. \quad (17)$$

T^*, or the final number of steps that are planned ahead, is now the value of T where the cost exceeds the gain in expected value from T to $T+1$.

The online model replicates the results derived from the offline version with myopic estimates as well as when retrospective experiences are incorporated (Figure 5). The absolute value of this model's planning estimates are larger, since the agent is receiving new prospective samples at each iteration. Additionally, we frame our results in terms of probability of expansion at each layer rather than planning depth.

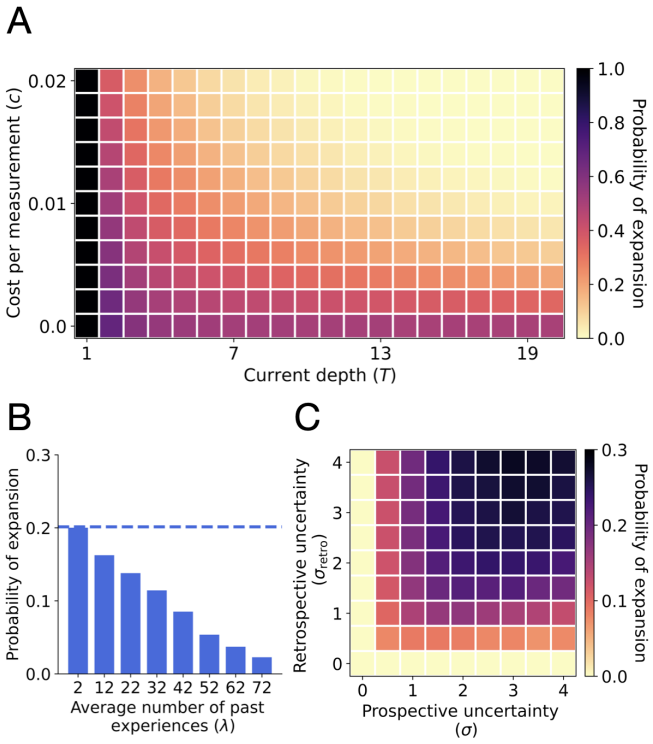

Figure 5: Online planning results for a fixed number of alternatives ($N=10$). **(A)** Probability of expanding to the next search iteration as a function of current depth (T) and cost per measurement (c). **(B)** Probability of expansion as a function of average number of retrospective measurements per action for a single cost ($c=0.01$). The dashed line indicates the probability of expansion without retrospective experience. **(C)** Probability of expansion as a function of retrospective and prospective variance in the generative model for the same cost as in (B).

Discussion

In this paper, we presented a normative framework for optimizing the depth of decision tree search. The model is based on using Bayesian inference to compute the value of planning from a combination of retrospective samples and myopic estimates. We began by showing that this model makes reasonable depth predictions without retrospective experience, primarily driven by cost per measurement and the number of actions. We also explained how the effect of the gap size between the top two myopic evaluations is in line with these results. We then introduced retrospective experience, and found that planning depth decreases as the agent gains experience and increases with the uncertainty of the model's evaluations unless prospective uncertainty is so high that deeper planning is no longer worthwhile. Finally, we derived an online version of the model and showed results consistent with its offline counterpart.

We must also evaluate how this framework might interact with planning in real tasks. Since we have conceptualized this as a metacognitive algorithm, the most natural extension is that the agent uses this approximation to quickly learn useful heuristics prior to planning. This heuristic can be a simple depth estimate given by the offline variant of the model prior to planning, or more nuanced components of a planning algorithm, such as a stopping rule, given by the iterative, online variant of the model. Another option is that this framework is actually implemented by the brain, in which case task-specific features and structure need to be incorporated into the model's evaluations. In either case, the model must be adapted to well-characterized planning tasks to verify that previous conclusions in the field, such as uncertainty arbitra-

tion and pruning, hold (Daw et al., 2011; Huys et al., 2012). Further, a primary motivation for developing this algorithm is to explain human behavior in more complex tasks with large data sets (Kuperwajs, van Opheusden, & Ma, 2019).

The method by which our model approximates the effects of search bears resemblance to the information sampling literature. In multi-armed bandit problems, people must choose between a set of alternatives that each have unknown reward in order to maximize total expected reward. Bayesian analyses of bandit problems exist, but typically provide a closed-form solution and focus on the tradeoff between exploration and exploitation (Steyvers, Lee, & Wagenmakers, 2009). In contrast, our framework focuses solely on determining high-value decisions. More recently, related work has claimed that simple decisions are made by integrating noisy evidence that is sampled over time in a Bayesian manner (Callaway, Rangel, & Griffiths, 2021; Jang, Sharma, & Drugowitsch, 2021). Our framework can be viewed as an approximation to planning via an optimal information sampling algorithm, and shares many features with these models. Conceptually, the main difference is in domain application, as prior work has explained fixation data in choice tasks with few alternatives while our model aims to derive intuitive rules to guide sequential decision-making. This is particularly relevant to the form that our model will take when interacting with a forward search algorithm in complex planning tasks.

While we have presented two model variants here, a more sophisticated variant of our model would determine online which action to expand the search frontier for, thereby optimizing both the depth and direction in which planning is beneficial. This can be thought of as best-first search, which combines breadth and depth by expanding the most promising node of the tree at each iteration. In future work, we plan to distinguish between how all three of these variants interact with human behavior in planning tasks.

Acknowledgments

This work was supported by Graduate Research Fellowship number DGE183930 and grant number IIS-1344256 from the National Science Foundation.

References

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018). A resource-rational analysis of human planning. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, *17*(3), e1008863.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67-81.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704-1711.

Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Pfaff, T., Weber, T., Buesing, L., & Battaglia, P. W. (2019). Combining q-learning and search with amortized value estimates. *arXiv*.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098-3103.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, *10*, e63436.

Kool, W., Gershman, S. J., & Cushman, F. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321-1333.

Kuperwajs, I., van Opheusden, B., & Ma, W. J. (2019). Prospective planning and retrospective learning in a large-scale combinatorial game. *Cognitive Computational Neuroscience*.

Sezener, C. A., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and direction of prospective planning using information values. *PLOS Computational Biology*.

Shannon, C. E. (1950). Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(314), 256-275.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Snider, J., Lee, D., Poizner, H., & Gepshtein, S. (2015). Prospective optimization with limited resources. *PLoS Comput Biol*, *11*(9), e1004501.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature neuroscience*, *22*(9), 1503–1511.

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Li, Y., & Ma, W. J. (2021). Revealing the impact of expertise on human planning with a two-player board game. *PsyArXiv*.