

# A way around the exploration-exploitation dilemma.

Erik J Peterson<sup>a,1</sup> and Timothy D Verstynen<sup>a,b</sup>

<sup>a</sup>Department of Psychology; <sup>b</sup>Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh PA

The exploration-exploitation dilemma is a fundamental but intractable problem in the learning and decision sciences. Here we challenge the common view of the dilemma—which focuses only on reward—by breaking into a tractable two-part problem, creating separate mathematical objectives for exploration and exploitation. To value exploration independently of reward, we develop a new set of values axioms inspired by Shannon’s original approach to developing information theory. We then turn to field of computer science, and the method of dynamic programming, to develop a provably optimal (greedy) algorithms that maximize both information value and reward. The cost of this solution is an increase in worst-case sample efficiency when compared to a equivalent reinforcement learning problem. The severity of the cost depends on the size of the environment and the rate of rewards.

When placed in a new environment a rat will explore it even if no tangible rewards, like food and water, are expected ( ? ? ? ). Sometimes they’ll explore even when it is certain that exploration will lead to a loss ( ? ). If a reward is expected though, an animal will generally explore and discover it ( ? ). This exploration gets interpreted as the animal exploring *for* reward—to maximize reward ( ? ). Here we offer an alternative view: animals don’t explore to maximize reward at all, they explore to maximize information—sometimes—about reward.

Our contribution is threefold. We first set out a series of axioms to serve as principled basis for information value, and use these to implement a working scheme rooted in information theory. Next we develop an optimal policy for exploration that maximizes information value. This theory is purely deterministic, so exploration no longer relies on random sampling ( ? ? ? ? ? ). Finally we derive a new myopic controller ( ? ), a *meta-policy*, that allows for simultaneously optimal and deterministic solutions to both problems of exploration and exploitation.

Exploration is almost universally regarded as stochastic process ( ? ? ). Before moving to stochastic view, we considered a deterministic approach. We’ll show how to recast exploration as a deterministic search process. Specifically, we’ll borrow a tool from computer science and cast exploration as a dynamic programming problem ( ? ). To do this we need to make an additional assumption about animal learning.

**A Bellman solution.** If  $M$  has optimal substructure, we can use the Bellman equation ( ? ) to write down an optimal (greedy) policy for maximizing information value  $E$  over the time horizon  $T$  is given by Eq. 1.

$$V_{\pi_E}^*(M_0) = \max_{a \in A} \sum_{t \in T} F(M_t, a_t) \quad [1]$$

By theorem 1, memory  $M$  has optimal substructure which

means Eq. 1 can be decomposed into a series of local greedy decisions.

$$\begin{aligned} V_{\pi_E}^*(M_0) &= \max_{a \in A} \left[ \sum_{t \in T} F(M_t, a_t) \right] \\ &= \max_{a \in A} \left[ F(M_0, a_0) + \sum_{t \in T} F(M_{t+1}, a_{t+1}) \right] \\ &= F(M_0, a_0) + \max_{a \in A} \left[ \sum_{t \in T} F(M_{t+1}, a_{t+1}) \right] \\ &= F(M_0, a_0) + V_{\pi_E}^*(M_{t+1}) + V_{\pi_E}^*(M_{t+2}), \dots \end{aligned} \quad [2]$$

From the final entry in Eq. 2, we can write down an optimal recursive value function for the current memory  $M_0$  in terms of the next memory  $M_1$  (Eq. 3).

$$V_{\pi_E}^*(M_t) = F(M_t, a_t) + \max_{a \in A} [F(M_{t+1}, a_t)] \quad [3]$$

Theorem 1 and Eq 2 demonstrate how to optimally maximize information value, given an optimal initial payoff  $F(M_0, a_0)$ . The question then is how find this first value optimally; A key step needed for all optimal dynamic programming solutions. To do this we imagine we have a vector of  $K$  possible payoffs  $F_0 \in \mathbb{R}^K$ , which is the same size as our action space  $A$ . If all the values of  $F_0$  are equal, then under a greedy policy any choice is a good as any other—so any choice is optimal. We restrict the initial value to  $F_0 > 0$ .

**Exploration as a dynamic programming problem.** The dynamic programming solution to maximize information value is also a dynamic programming solution for exploration. By making a weak assumption on the learning of  $M$ , we can prove that our greedy policy  $\pi_E^*$  leads to a complete and exhaustive exploration of any finite space  $S$ . Specifically, we show:

1. The optimal policy  $\pi_E^*$  must visit each state in  $s \in S$  at least once.
2. The optimal policy  $\pi_E^*$  must revisit each  $s \in S$  until learning about each state  $s$  plateaus when  $g(M_{t+1}, s) = g(M_t, s)$ .

**Exploration during reinforcement learning.** The overall objective in reinforcement learning is to maximize  $V_{\pi_R}$ , the expected value of total future rewards (Eq 4) ( ? ). Here we use the term  $I_t$  as a generic stand in for a range of fictive rewards, including novelty bonuses ( ? ), curiosity signals ( ? ), Bayesian updates ( ? ), or entropy terms ( ? ? ). Often the fictive term in these equations is arbitrarily re-weighted. We use  $\beta$  for this fictive weight. Likewise, rewards are denoted by  $R_t$ . In this

The authors have no conflicts of interest to declare.

<sup>1</sup>To whom correspondence should be addressed. E-mail: Erik.Exists@gmail.com

example the total time  $T$  is said to be finite, simplifying the notation.

$$V_{\pi_R} = \sum_{t \in T, s \in S} R_t + \beta I_t \quad [4]$$

Explicitly, we propose that reinforcement learning problems should try and optimize two values: reward (Eq. 5) and information (Eq. 6).

$$V_{\pi_R} = \sum_{t \in T, s \in S} R_t \quad [5] \quad V_{\pi_I} = \sum_{t \in T, s \in S} I_t \quad [6]$$

We separate the objectives because classical rewards and information value have fundamentally different philosophical and mathematical properties. Conflating them introduces an undesirable bias, both in trying to maximize information value and in maximizing rewards.

**Biased by reward.** Exploration often means a loss of rewards. To motivate exploration, lost rewards are often re-balanced with information gains, novelty signals, or other fictive reward signals. This is simple and intuitive, but comes with an under-appreciated limitation. If the true goal is to maximize total reward  $R$ , adding a fictive reward obscures this aim. After all, now an increase in  $V_{\pi_R}$  may come from  $R$  or from the fictive term  $\beta I$ . So while fictive rewards do encourage exploration they offer no guarantee of a performance increase in  $R$ , or in the long-term value  $V_{\pi_R}$ .

Conversely, if an animal wants to build a generally useful model of world, as Schmidhuber suggests, then biasing choices by maximizing reward is suboptimal. Both in terms of maximizing both the accuracy of the world model (which will wind up over-sampled near rewarding sites), and the efficiency with which the model is built (much time is wasted gathering rewards).

**Information is not a reward.** Information and reward so seem to have opposing properties. If a rat shares a potato chip with a cage-mate, it must break the chip up to share, leaving has less food for itself. While if a student shares an idea with a lab-mate, the idea is not divided up or lost. Thus rewards are a conserved resource, information is not.

The same kind of reward can be consumed many times without necessarily losing value\*. Information once learned though has no value in being learned again. Eating one potato chip often means wanting another, whereas if you know the capital of the United States, there is no value in being told the capital of the United States is Washington DC.

These two philosophical differences suggest that reward and information may also have notable mathematical differences. To evenly share a reward  $r$  between  $n$  others,  $r_n = \frac{r}{n}$ . In contrast, information value can, in principle, be shared without loss, i.e.,  $I_n = I$ . Likewise, if an animal is learning about some state of world  $s$  the value of information about  $s$  should decrease with time. Assuming learning can asymptote, then as  $t \rightarrow \infty$ ,  $I_s \rightarrow 0$ . For rewards, value never changes. So as  $t \rightarrow \infty$ ,  $r_s \rightarrow r_s$ .† In summary, information is fundamentally

\* In practice, there are satiety effects but these are not a necessary part of the reward's value

† Though including reward satiety effects may somewhat diminish  $r_s$  in practice.

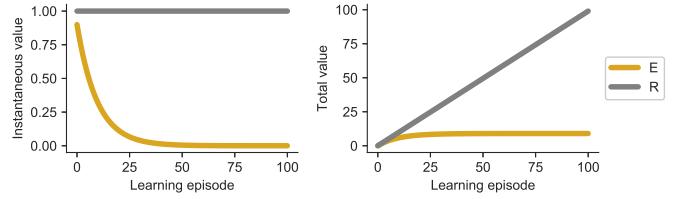


Fig. 1. Simulated learning of information and reward value, assuming a linear reward accrual and an exponential decay information value. **a.** Instantaneous value over 100 learning epochs. **b.** Average value over 100 learning epochs.

not a reward. Information about the world and the rewards from the world are very different kinds of things.

**A way around the dilemma.** We introduce now a new style of reinforcement learning we call *dual value learning*. We conjecture that from point of the view of an animal learning to explore its world *and* gather the most rewards, it can, and probably should, use two different policies for each aim.

Having two policies  $\pi_E$  and  $\pi_R$  sidesteps the exploration-exploitation dilemma in the following ways. Exploration now generates tangible value, and so is not in any sense costly. Because of this exploration can be carried out completely. In turn this ensures the reward policy  $\pi_R^*$  discovers its global optimum, if it exists. In circumstances where the environment can be learned  $E$  will decay to  $\epsilon \geq 0$  leaving  $\pi_R^*$  in control, and so exploitation becomes the final policy. When the environment changes exploration can begin again, as  $\pi_E^*$  resumes intermittent control. Finally, while only one policy can control behavior each policy can learn from the actions of the other. This makes learning to maximize exploration and exploit rewards cooperative, even though action control is competitive.

One policy can drive behavior at a time. To adjudicate, we derive a simple, value-based, myopic controller (?) to select which policy dominates. We call this policy on policies a *meta-policy* and denote it by  $\pi_\pi$ . We insist that for any meta-policy to be considered valid meta-policy it must inherit any optimality present in  $\pi_E^*$  and  $\pi_R^*$ .

We base our formulation for  $\pi_\pi$  in Eq 7 on the asymptotic properties of information and rewards. To review,  $E_t \rightarrow \epsilon$  as  $t \rightarrow T$  (Theorem 2) which implies that as  $T \rightarrow \infty$  then  $V_{\pi_E} \rightarrow \epsilon$ . Reward, on the other hand does not decay with time and so  $V_{\pi_R}$  can grow without bound. A cartoon example of this is depicted in Figure 1.

$$\pi_\pi = \begin{cases} \pi_E & : E_t - \epsilon > R_t \\ \pi_R & : R_t \geq E_t - \epsilon \end{cases} \quad [7]$$

subject to the constraints

$$\begin{aligned} R_t &\in \{0, 1\} \\ p(R_t = 1) &< 1 \\ E_t - \epsilon &> 0 \end{aligned}$$

Note that to satisfy  $E_t - \epsilon > 0$  in Eq 7,  $\epsilon = 0$  when  $\Lambda(\cdot)$  is deterministic. Only when the environment is stochastic can  $\epsilon$  safely exceed 0.

The meta-policy  $\pi_\pi$  is a myopic in the sense that at every time  $t$ ,  $\pi_E$  or  $\pi_R$  has the opportunity to take control but this control lasts only a single time step. There are many other

possible meta-policies which could also inherit the optimality of our dual policies. We chose this form for its simplicity, but justify it as similar myopic controls have proven effective for other complex, high variance, problems in neuroscience (? ).

Even though the individual policies in the meta-policy are independent, learning can be done in parallel. Regardless of which policy is in control, they can in principle observe choices made by the other and learn from them.

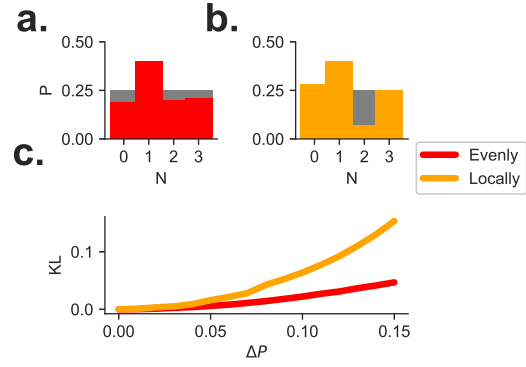
**Theorem 1** (Optimality of  $\pi_\pi$ ). *Assuming an infinite time horizon, if  $\pi_E$  is optimal and  $\pi_R$  is optimal, then  $\pi_\pi$  is also optimal in the same sense as  $\pi_E$  and  $\pi_R$ .*

*Proof.* The optimality of  $\pi_\pi$  can be seen by direct inspection. If,  $p(R = 1) < 1$  and we have an infinite horizon, the  $\pi_E$  will have a unbounded number of trials meaning the optimality of  $P^*$  holds. Likewise,  $\sum E < \epsilon$  as  $T \rightarrow \infty$ , ensuring  $p_{iR}$  will dominate  $\pi_\pi$  therefore  $\pi_R$  will asymptotically converge to optimal behavior.  $\square$

In proving the total optimality of  $\pi_\pi$  we limit the probability of a positive reward to less than one, denoted by  $p(R_t = 1) < 1$ . Without this constraint the reward policy  $\pi_R$  would always dominate  $\pi_\pi$  when rewards are certain. While this might be useful in some circumstances, from the point of view  $\pi_E$  it is extremely suboptimal as the model would never explore. Limiting  $p(R_t = 1) < 1$  is reasonable constraint, as rewards in the real world are rarely certain. A more naturalistic but complex way to handle this edge case would be to introduce reward satiety, and have reward value decay asymptotically with repeated exposure.

**An optimally rewarding policy.** If learning during  $\pi_\pi$  is allowed to continue until  $\pi_E$  converges so  $E_t - \epsilon = 0$  then, by definition the animal will have completely explored its world. This implies that in turn  $\pi_R$  has seen every state, and so can then choose the overall optimal value. Thus is there is a globally optimally reward policy,  $\pi_\pi$  guarantees it will be found. Classic reinforcement learning views this search as costing potential reward. Dual value learning instead asserts a net gain. Either from information value, from rewards, or both. We suggest that rather than being fundamental, the exploration-exploitation dilemma follows from asking too little from an animal's learning objectives.

**E is satisfied by KL.** The Kullback–Leibler divergence (KL) is a widely used information theory metric, which measures the information gained by replacing one distribution with another. It is highly versatile and widely used in machine learning (? ), Bayesian reasoning (? ? ), visual neuroscience (? ), experimental design (? ), compression (? ? ) and information geometry (? ), to name a few examples. Using a Bayesian approach, Itti and Baladi (? ) developed an approach similar to ours for visual attention, where our information value is identical to their *Bayesian surprise*. Itti and Baladi (2009) showed that compared to range of other theoretical alternative, information value most strongly correlates with eye movements made when humans look at natural images. Again in a Bayesian context, KL plays a key role in guiding *active inference*, a mode of theory where the dogmatic central aim of neural systems is make decisions which minimize (probabilistic) free energy (? ? ).



**Fig. 2.** Local probability structure and information value. Both distributions shown in a. and b. have the same total increase in the probability of a “1” appearing. For a. the necessary corresponding decrease in probabilities for all numbers (0, 2, 3) is evenly redistributed. In b. the loss is focused locally on “2”. c. The KL divergence increases more rapidly for local changes (orange) in probability density compared to an even re-distribution of probability mass (red)

The Kullback–Leibler (KL) divergence satisfies all five value axioms (Eq. 8). In expressing  $E$  in terms of KL it also allows us to more concretely demonstrate the mathematical properties implied in our axioms.

$$KL(M', M) = \sum_{s \in S} M'(s) \log \frac{M'(s)}{M(s)} \quad [8]$$

**Definition 1.** Let  $E$  represent value of information, such that  $E = KL(M', M)$  (Eq. 8), where  $M$  is some initial memory and  $M'$  is an update memory after observing some state  $s'$ .

Axiom 1 is satisfied by limiting  $E$  calculations to successive memories. Axiom 2-3 are naturally satisfied by KL. That is,  $E = 0$  if and only if  $M' = M$  and  $E \geq 0$  for all pairs  $(M, M')$ .

To make Axiom 5 more intuitive, in Figure 2 we show how KL changes between an initial distribution (always shown in grey) and a “learned” distribution (colored). For simplicity’s sake we use a simple discrete distribution, representing the likelihood of observing the first four integers, (0, 1, 2, 3). Though the illustrated patterns should hold true for any pair of distributions. In Figure 2 we see KL increases substantially more, for a the same local increase in probability, when that increase comes with a localized decrease, rather than with an even re-normalization (compare panels a. and b.).

That is, in Figure 2c we explore how KL changes with a change in total uncertainty,  $\Delta P$ . In panels a. and b. the (uniform) grey distribution represents our baseline. The colored distribution represents the largest change in uncertainty, corresponding to  $\Delta P = 0.15$ . Along the x-axis we see how KL responds to increases in  $\Delta P$ , depending on whether that is  $\Delta P$  is distributed evenly (a) or more specifically (b). In both cases though, and in line with Axiom 4, we can also see how KL increases monotonically with  $\Delta P$ .

**Limitations.** A deterministic  $\pi_E$  requires that the initial set of information values,  $E_0$ , be provided. If  $E_0 = 0$ , the meta-policy will never begin any exploration. While if  $E_0 > 0$ , theorem 1 ensures that any  $E_0$ ,  $\pi_E$  will optimal in terms of maximization of total  $E$ . Likewise, theorem 2 ensures that any initial choice will not effect the optimality of the search. The magnitude of  $E_0$  does not change  $\pi_E$ ’s long term behavior,

but will of course change ins transient dynamics which might be important when applying this work to real life settings.

In our simulations we assume that at the start of learning an animal should have a uniform prior over the possible actions  $A \in \mathbb{R}^K$ . Thus  $p(a_k) = 1/K$  for all  $a_k \in A$ . We transform this uniform prior into the appropriate units for our KL-based  $E$  using Shannon entropy,  $E_0 = \sum_K p(a_k) \log p(a_k)$ .

By definition a greedy policy can't handle ties, as there is no single way to rank equal values. Our theorems 3 and 2 ensure that any tie breaking strategy is valid. However like the choice of  $E_0$ , tie breaking can strongly effect the transient dynamics of  $\pi_E$  and so can be quite important in practice. Viable tie breaking strategies taken from experimental work include, "take the closest option", "repeat the last option", or "take the option with the highest marginal likelihood". We do suggest the tie breaking scheme is deterministic, which maintains the determinism of the whole theory. In our simulations we use a tie breaking heuristic which keeps track of past breaks, and in a round robin fashion iterates over the action space.

Optimal exploration, both in terms of search and max  $E$  does not promise that the agent has learned an accurate or unbiased memory. The loss function  $\mathcal{L}_M$  is responsible for that. By comparing a memory before and after learning, information value only measures self-consistency. That is, the agent could learn a bad model of the world, but as long as it learns it consistently exploration will be consider convergent and total  $E$  optimal.

Previous work studying information gain and memory has adopted an explicitly Bayesian view of animal cognition, as seen for example in (??). Bayesian cognition is strongly normative account, which is mechanistically and experimentally controversial (?). Bayesian cognition is not an assumption we require. Instead we study exploration as a simple optimization problem, whose best policy is expressed by dynamic programming.

**Non-convex learning.** To simplify the problem we have assumed that the memory learning loss function  $\mathcal{L}_M$  is both convex, and its gradient is always negative. Intuitively, these assumptions mean 1. there is a global solution to learning  $M$  and that 2. with every observation some small amount of learning progress is made. In practice, neither assumption is realistic. Many of our most powerful learning systems are non-convex, and not all observation can and do lead to learning progress.

There are many reasons learning can diverge. For example, a "bad" initialization or "bad" hyper-parameters, too much noise in the input data, and so on. For many complex and realistic environments, convergence of any given learning algorithm, even a convex one, is not guaranteed in the general case (?). If learning does not at least begin to converge, then  $E$  will not converge and  $\pi_E$  must be non-optimal. This is, however, a fundamental limitation of learning theory. Still, our analysis must confine itself to agents that do learn.

If learning begins to diverge so  $\nabla \mathcal{L}_M > 0$ , then  $E$  must also grow (see Eq. 8). If noise or stimulus complexity drive this momentary loss of learning progress, increases to  $E$  will force the animal to "resample" these stimuli. The resampling order is based on the expected information gain with each. In many circumstances we conjecture maximizing  $E$  will lead to  $\mathcal{L}_M$  both to resume a negative trajectory and do so as efficiently or quickly as possible. In this case, maximizing  $E$  remains a sound objective. On the other hand, if  $\mathcal{L}_M$  diverged due a

problem with the learning algorithm itself, say a bad seed in a deep reinforcement learning model, then maximizing  $E$  may exacerbate the problem leading to a potentially catastrophic feedback loop. In this case maximizing  $E$  seems unsound, though it is not clear if any action policy would be better.

**Leaving local minimum.** Local minimum are a common concrete case of that any purportedly general theory of exploration must accommodate. We've assumed  $\nabla \mathcal{L}_M < 0$ . We can invert this assumption to extend our Theorems 1 and 2 to local minimum.

Assume instead  $\nabla \mathcal{L}_M > 0$  but this is constrained to a finite duration  $W$ , over an a sampling time  $T$ . A finite divergence like the above is half of a working definition for local minimum in  $\mathcal{L}_M$ . The other half being the minimum itself, defined by the gradient finding zero as  $\nabla \mathcal{L}_M = 0$ .

**Theorem 2 (Local minimum).** *For some learning time  $T$ , assuming the period  $W$  for which  $\nabla \mathcal{L}_M > 0$  is finite, then the known optimal policy  $\pi_E$  (per Theorem 2 and 3) is still optimal.*

*Proof.* If for the finite period  $W$ ,  $\nabla \mathcal{L}_M > 0$ , it's logically required also that at all other times  $T - W$ ,  $\nabla \mathcal{L}_M \leq 0$ . If  $\nabla \mathcal{L}_M = 0$  and  $E_t = 0$  then there is a tie, which can be broken arbitrarily per Theorem 2. Now, if hold  $W$  to be finite but let  $T \rightarrow \infty$  then the ratio  $\frac{T-W}{W}$  also must approach  $\infty$ . Thus by asymptotic analysis, we prove that whatever happens during  $W$  will be dominated by the known optimal period, as  $T \rightarrow \infty$  then  $\frac{T-W}{W} \rightarrow \infty$ .  $\square$

As we note above, there is no way in the general case to prove that any minimum is local, or that  $\mathcal{L}_M$  will overcome any local minimum. We show here that if a minimum can be overcome, greedily maximizing  $E$  will not interfere. We also conjecture in many cases maximizing  $E$  may be the most efficient re-sampling strategy.

**Simulating behavior.** *Note: simulations are work in progress. . .*

### Properties of the meta-policy.

**Sample efficiency.** The meta-policy  $\pi_\pi$  is form a myopic control where only one of the two dual policies,  $\pi_E^*$  or  $\pi_R^*$ , can control action selection at time. So if we define the number of state observation as the number of samples, the worst case sample efficiency for  $\pi_\pi$  is additive in its policies. That is, if in isolation it takes  $T_E$  steps to earn  $E_T = \sum_{T_E} E$ , and  $T_R$  steps to earn  $r_T = \sum_{T_R} R$ , then the worst case training time for  $\pi_\pi$  is  $T_E + T_R$ . There is however no reason each policy can't observe the transitions  $(s_t, a_t, R, s_{t+1})$  caused by the other. If this kind of parallel learning is allowed, the worst case training time improves substantially to  $\max(T_E, T_R)$ . That is, learning can be done in parallel-making it cooperative-but action control is adversarial, governed by our myopic inequality  $\pi_\pi$  (Eq. 7).

**Exploration-exploitation as an initial value problem.** In our analysis  $\pi_E$  has been assumed to be deterministic. If we also restrict  $\pi_R$  to be deterministic-which is sensible when optimal exploration is certain-we can the define exploration-exploitation dilemma strictly as an initial value problem, which has at least one major benefit.



Exact, turn by turn, fits are of real behavior are possible. The experimenter need only hypothesize about initial conditions. This means specifying the initial value of  $E_0$  (i.e., its prior) and establishing a tie breaking rule, as well as setting the hyper-parameters. At minimum hyper-parameter tuning will require setting the learning rates for both policies in  $\pi_\pi$  (Eq 7), and the convergence threshold for exploration,  $\epsilon$ . Critically though there is no need to hand tune sampling noise (?).

**The rates of exploration and exploitation.** In Theorem 1 we proved that  $\pi_\pi$  inherits the optimality of policies for both exploration  $\pi_E$  and exploitation  $\pi_R$  over infinite time. However this does proof does not say whether  $\pi_\pi$  will not alter the rate of convergence of each policy. By design, it does alter the rate of each, favoring  $\pi_R$ . As you can see in Eq. 7, whenever  $r_t = 1$  then  $\pi_R$  dominates that turn. Therefore the more likely  $p(r = 1)$ , the more likely  $\pi_R$  will have control. This doesn't of course change the eventual convergence of  $\pi_E$ , just delays it in direct proportion to the average rate of reward. In total, these dynamics mean that in the common case where rewards are sparse but reliable, exploration is favored and can converge more quickly. As exploration converges, so does the optimal solution to maximizing rewards.

**Re-exploration.** The world often changes. Or in formal parlance, the world is non-stationary process. When the world does change, re-exploration becomes necessary. Tuning the size of  $\epsilon$  in  $\pi_\pi$  (Eq 7) tunes the threshold for re-exploration. That is, once the  $\pi_E^*$  has converged and so  $\pi_R^*$  fully dominates  $\pi_\pi$ , if  $\epsilon$  is small then small changes in the world will allow  $\pi_E$  to exert control. If instead  $\epsilon$  is large, then large changes in the world are needed. That is,  $\epsilon$  acts a hyper-parameter controlling how quickly rewarding behavior will dominate, and easy it is to let exploratory behavior resurface.

**Extensions to the meta-policy.** Animals in the real world exhibit a large repertoire of behavior than simply exploration and exploitation. Without adding any new value terms, a range of other naturalistic behaviors can be mixed into our meta-policy approach.

**Aversion.** Recognizing how important aversive learning was to survival in both real and artificial agents, Schmidhuber (?) incorporated this in his formulation. We've meanwhile focused on reward learning in our analysis. Here we show how out myopic controller, the meta-policy, can be easily expanded to included aversive learning. When the last outcome was aversive, the animal may wish to follow a stimulus avoidance policy  $\pi_A$  on the next times step. If we let  $R = -1$  code for such aversive events, it is straightforward to incorporate this into a new meta-policy (Eq 9).

$$\pi_\pi = \begin{cases} \pi_A & : R_t < 0 \\ \pi_E & : E_t - \epsilon > R_t \\ \pi_R & : R_t \geq E_t - \epsilon \end{cases}$$

[9]

subject to the constraints

$$\begin{aligned} R_t &\in \{-1, 0, 1\} \\ p(R_t = 1) &< 1 \\ E_t - \epsilon &> 0 \end{aligned}$$

**What do to by default.** If overall stimulation / motivation is too low, an animal will stop exploring or exploiting, often instead adopting some default action policy  $\pi_\emptyset$  (e.g., grooming or checking Facebook). As a working example of this kind of meta-policy see Eq. 10. Here, when the average information  $\bar{E}$  and reward  $\bar{R}$  fall below the exploration threshold,  $\pi_\emptyset$  then dominates

$$\pi_\pi = \begin{cases} \pi_\emptyset & : (\bar{E} + \bar{R}) < \epsilon \\ \pi_E & : E_t - \epsilon > R_t \\ \pi_R & : R_t \geq E_t - \epsilon \end{cases}$$

[10]

subject to the constraints

$$\begin{aligned} R_t &\in \{0, 1\} \\ p(R_t = 1) &< 1 \\ E_t - \epsilon &> 0 \end{aligned}$$

## Discussion

Our analysis raises the possibility that reinforcement learning on just rewards is an incomplete theory. This has been partially acknowledged by adapting information terms (as fictive rewards) into reinforcement learning problems. However we suggest that on empirical, computational, mathematical, and philosophical grounds the use fictive rewards is not far enough. Instead, we believe information should be valued entirely for its own sake, and reward and information should be maximized separately. While this approach might seem more complex, we offer a simple optimal meta-policy for doing dual optimization. If an animal were to use this meta-policy, the classic exploration-exploitation dilemma is no longer a dilemma, and exploration can be both optimal and deterministic.

When training animals to learn a new task one of the most difficult parts in the training is constraining behavior. Left to their own devices and motivations, animals engage in a range of task-irrelevant activities. This kind of exploration is a nuisance to the experimenter, but we suggest is a important object of study for the theorist. Dual value learning accommodates the analysis and prediction for any mode of free ranging behavior, given a working definition for the state space  $S$ , a memory  $M$ , and the memory learning rule  $J$ .

**Curiosity and value.** A skeptical reader might suggest we've arbitrarily swapped an accepted term *curiosity*, for another *information value*. We introduce a new term for two reasons. 1. When separating information value from reward we felt it was important to place information value on principled grounds. This is why we developed the value axioms. Applying these axioms to an existing, and loosely defined term, like curiosity would seem to add confusion to the literature rather than simplify. 2., curiosity as a subjective experience seems to dim more quickly than learning a detailed model requires. That is, learning needs more than just curiosity. We try and capture this admittedly complex motivation using information gain and the KL divergence.

**On the axioms.** Given that KL measures the information gain (or loss) between two models, and has a long and useful history we could have skipped the Axiomatic approach and made a direct argument for KL. We did not do this for two reasons.

First, the value of a reward is based on its biological significance. Food, water, mates, are necessary for survival. If we were to value information for its own sake we felt it was critical to base that value not on a particular theoretical view (i.e. information theory) but instead on a firm set of theory-independent but well motivated principles. We felt obliged to first answer the question, “If information is valuable, what do we base that value on?”. We’ve made one answer to that question with our axioms; We hope though ours isn’t the last word.

Second, though (Shannon) information theory and KL are very useful constructs, they require symbolic and probabilistic representations. We don’t know whether animals actually maintain such exact representations. Likewise, in machine learning memory modules often don’t rely on distributions, and instead simple recall selections of previous events (For example, (?)). Our axioms can be satisfied in these kinds of cases. Likewise, our key Theorems 1-1 either don’t explicitly depend on information theory and KL, or can be trivially adapted to other axiomatic cases.

**On Axiom 5.** A skeptical reader might also find Axiom 5 to be overly opinionated. A simpler natural set of Axioms is to keep 1-3 but replace 5 with a new Axiom that only requires information value track the total change in probability flux, that is  $E \sim dP$ . For this metric, the curves in Figure 2 would overlap. We considered this, but felt this simpler view is incomplete. Information value should favor more specific, therefore more actionable, information.

**Learning structure and value.** Here we’ve explored only a simple probabilistic memory model that lends itself to information theoretic calculations. There are a large number of approaches an animal might use to build a model of the world. These include Bayesian structure learning, dynamical systems modeling, casual reasoning, By defining information value axiomatically, we offer a principled and consistent way to fold potentially any model-build method seamlessly and optimally into a value optimization framework. Doing this might require developing a new metric other than the KL divergence. We conjecture though that in many cases adding a simple probabilistic memory module to estimate state-action likelihoods—like the one we study here—may prove sufficient.

16. Hocker D, Memming I, Brook S (2019) Myopic control of neural dynamics. pp. 1–30.
17. Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. (MIT Press).
18. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. *Vision Research* 49(10):1295–1306.
19. Friston K, et al. (2016) Active inference and learning. *Neuroscience and Biobehavioral Reviews* 68:862–879.
20. López-Fidalgo J, Tommasi C, Trandafir PC (2007) An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 69(2):231–242.
21. Mackay DJC (2003) *Information Theory, Inference, and Learning Algorithms*. (Cambridge university press).
22. Still S, Sivak DA, Bell AJ, Crooks GE (2012) Thermodynamics of prediction. *Physical Review Letters* 109(12):1–5.
23. Ay N (2015) Information geometry on complexity and stochastic interaction. *Entropy* 17(4):2432–2458.
24. Sutton RS, Barto A (2018) *Reinforcement Learning*. (MIT Press).
25. Schmidhuber J (1991) A possibility for implementing curiosity and boredom in model-building neural controllers in *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats*. Vol. 1, pp. 222–227.
26. Min HK, et al. (2016) Dopamine Release in the Nonhuman Primate Caudate and Putamen Depends upon Site of Stimulation in the Subthalamic Nucleus. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 36(22):6022–9.
27. Shannon C (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal* XXVIII(3).
28. Kolchinsky A, Wolpert DH (2018) Semantic information, autonomous agency and non-equilibrium statistical physics.
29. Deacon TW (2015) Shannon - Boltzmann — Darwin : Redefining information ( Part II ). 2(2):169–196.
30. Tishby N, Pereira FC, Bialek W (year?) The Information Bottleneck Method. pp. 1–11.

1. Liu A, et al. (2019) Mouse navigation strategies for odor source localization. *BioRxiv* 558643.
2. Jaegle A, Mehrpour V, Rust N (year?) Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. pp. 1–13.
3. Todd PM, et al. (2015) Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* 2(3):191–215.
4. Zhe Wang M, Hayden BY (2019) Monkeys are curious about counterfactual outcomes.
5. Sutton RS, Barto A (2018) *Reinforcement Learning: An Introduction*. (MIT Press), 2 edition.
6. Kakade S, Dayan P (2002) Dopamine: Generalization and bonuses. *Neural Networks* 15(4-6):549–559.
7. Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*.
8. Sutton RS (1990) Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. *Machine Learning Proceedings 1990* 02254(1987):216–224.
9. Thrun SB, Möller K (1992) Active exploration in dynamic environments in *Advances in neural information processing systems*. pp. 531–538.
10. Dayan P, Sejnowski TJ (1996) Exploration bonuses and dual control. *Machine Learning* 25(1):5–22.
11. Hocker D, Memming I, Brook S (2017) Myopic control of neural dynamics. pp. 1–24.
12. Bellman R (1957) *Dynamic Programming*. (Princeton University Press).
13. Radulescu A, Niv Y, Ballard I (2019) Holistic Reinforcement Learning : The Role of Structure and Attention. *Trends in Cognitive Sciences* xx:1–15.
14. Haarnoja T, Zhu H, Tucker G, Abbeel P (2015) Soft Actor-Critic Algorithms and Applications.
15. Haarnoja T, Tang H, Abbeel P, Levine S (2017) Reinforcement Learning with Deep Energy-Based Policies.

## Mathematical appendix

**Value axioms.** Shannon developed information theory without any sense of what information “means”. He focused on transmitting symbols, not what those symbols refer to. This made the theory extremely general (? ). We seek a similarly general way to value information. To do this we followed Shannon’s lead and developed a set axioms. But unlike Shannon we don’t study how symbols in an alphabet are transmitted. We study how states from the world are remembered.

The basic idea is simple. The more memory changes when observing a state, the more valuable that observation is. It doesn’t matter what that state was, only how the memory changes.

Many attempts have since been since Shannon to instill information theory with meaning and value (? ). Most of these attempts require a “salience”, or “relevance”, or “training” signal (where all three terms amount to near the same thing; (? ? )), while others have used evolutionary theory to create a reference (? ? ). Here we require no outside reference to estimate value.

To state the axioms first we need to introduce some notation and definitions. We ask for a bit of patience from the reader. The definitions are necessarily abstract, as we are developing a theory of information value that is independent of any particular model of animal cognition or learning.

**States and actions.** Here we study a number of  $N$  finite states  $s$  in a world  $s \in S^N$ . This is analogous to Shannon’s initial formulation of symbols  $x$  in an alphabet  $X$ . A state might be a visual scene (? ), or an odor vector (? ? ), an index into a symbol table, or a set of spatial coordinates, or any set of numbers, vectors, or tensors. All that matters is the states are unique and self-consistent in  $S$  and that they do not change with time or with experience.

We limit the number of actions  $a$  animal might take to a finite set  $A$  of size  $K$ ,  $a \in A^K$ . Like  $S$ , we allow the set  $A$  to conceivably be any set of numbers, vectors, or tensors.

**A definition of memory.** A memory  $M$  is a set of transformed states, defined recursively by the encoder  $f$  as  $M \leftarrow f(M, s)$ . The base case for  $M$  is said to be the empty set  $M_0 = \emptyset$ . In defining  $f$  we assume the encoding of any  $s$  can, in principle, create or modify multiple memories in  $M$ . That is we do not assume a bijection and not only one element  $m_i \leftarrow f(s, m_i)$  from  $M = \{m_1, m_2, m_i, \dots\}$ . We do not require  $M$  be finite like  $S$  which means in theory it’s possible to form more memories than there are states.

We require that the memory  $M$  can decoded by function  $g$  to produce recollection  $z$  by  $g(M, s) \rightarrow z$ . We don’t concern ourselves with details of  $z$ . They are as open as the definition of states in  $S$ .

We define time  $t$  as a index of sequential state observations of  $s \in S$ , such that  $t = (0, 1, 2, 3, \dots, T)$ .

To finish our definition of  $M$  we add one critical restriction to the encoder  $f$ : it must be invertible. That is, if there exists  $M_{t+1} = f(M_t, s)$  then there must also exist an inversion  $f^{-1}$  so  $M_t = f^{-1}(M_{t+1}, s)$ .

We summarize the requirements for memory  $M$  in Definition 1.

**Definition 1.** We define a memory  $M$  as a potentially infinite set, defined over a finite state space  $s \in S^N$ , whose elements

are added by an invertible encoder  $f$  such that  $M_{t+1} = f(M_t, s)$  and  $M_t = f^{-1}(M_{t+1}, s)$ . Elements  $z$  from  $M$  are recovered by a decoder  $g$ , such that  $z = g(M, s)$ . The initial memory  $M_0$  is the empty set,  $M_0 = \emptyset$ .

Definition 1 covers simple and complex ideas of a memory. In the simplest case where  $f$  and  $g$  are identity functions, then  $M$  is just a simple direct record, or subset, of the states observed by the animal  $M \subseteq S$ . If  $f$  is a probability function,  $0 \leq f(s) \leq 1$ ;  $\sum_s f(s_i) = 1$ , then  $M$  is a distribution over  $S$ , and  $g(M, s_i) = p(s_i)$ . Likewise, if  $f$  is not only a probability function implements Bayes’ rule, then  $M$  can be a conditional distribution. In principle Definition 1 also encompass hierarchical Bayesian learning, VAE and other generational approaches found in modern machine learning / artificial neural networks, as well as sequential and recurrent memories, like echo state networks, FORCE, LSTM and GRU. It also extends also to cases not often considered memories *per se* but which include an encoder-decoder pair– including dimensionality reduction techniques (*e.g.* PCA, NMF, tSNE), clustering methods (KNN), and mixture models (GMM, kernel).

Several of these methods—especially those we’ve labeled as more complex—are defined implicitly or explicitly with an encoder and decoder. Most do not explicitly or implicitly include an inverse decoder  $f^{-1}$ . However all we require is that  $f^{-1}$  could be constructed, *in principle* (Theorem 1).

**Memory space.** As we’ll seen in the next section, to make information value axiomatic we rely on comparing sequential changes to the memory  $M$ , as it moves from  $M_t$  to  $M_{t+1}$ . This comparison is made using a distance metric  $d(x, y)$ , which is minimally constrained by  $d(x, y) \geq 0$  and  $d(x, y) = 0$  if and only if  $x = y$ . We do not require that  $d$  obeys the triangle inequality, nor that is be symmetric.

**Definition 2.** We define a distance  $d$  to be any function where  $d(x, y) \geq 0$  and  $d(x, y) = 0$  if and only if  $x = y$ .

The distances for two sequential memories  $M_t$  to  $M_{t+1}$  are measured using the distance  $d$  on the decoded set of memories  $d(z_{t+1}, z_t)$ . This assumption let’s us separate the definition (and eventual implementation) of  $M$  from the distance  $d$ .

**The value axioms.** We denote information value as  $E$ , subject to the following axiomatic constraints. Under each formal Axiom, we offer an informal summary.

**Axiom 1.**  $E$  is a function which measures the distance  $d$  between  $M_{t+1}$  and  $M_t$  to a real number,  $E : (M_{t+1}, M_t) \rightarrow \mathbb{R}$ .

Value is subjective, and so should only depend on an animals memory. (Working with just sequential memories,  $M_{t+1}$  and  $M_t$  and scalars, simplifies the problem mathematically.)

**Axiom 2.**  $E \geq 0$ .

New information about the world is always valuable, even if its later consequences might not be. The consequences of the information may not always be good, but that’s a separate issue.

**Axiom 3.**  $E = 0$  if and only if  $M_{t+1} = M_t$ .

If you remember something perfectly, there is no value in learning it again.

**Axiom 4.**  $E$  is strictly monotonic with the total change in  $d$  measured between  $M_{t+1}$  and  $M_t$ .

The information value should track the total change in memory.

**Axiom 5.**  $E$  is monotonic with the compactness  $C$  (Eq. 1) of  $d$  measured between  $M_{t+1}$  and  $M_t$ .

Specific information is more always more valuable than less specific information. (We use geometrical compactness as surrogate for specificity, which doesn't seem to have a precise meaning.)

The compactness  $C$  of a hyper-cube has a simple formula,  $C = \frac{P^2}{A}$  where  $P$  is the cube's perimeter and  $A$  is its area. Therefore as  $M_t$  moves to  $M_{t+1}$ , the we can measure the distances  $\{d_i, d_{i+1}, d_{i+2}, \dots, d_N\}$  for all  $0 \leq N$  observed states  $Z$ , such that  $Z \subseteq S$  and treat them as if they formed a  $O$ -dimensional hyper-cube. In measuring this imagined cube we arrive at a geometric estimate for change the compactness of our memory  $M$  (Eq. 1).

$$C = \frac{(2 \sum_O^i d_i)^2}{\prod_O^i d_i} \quad [1]$$

**Information value as a dynamic programming problem.** To use theorems from dynamic programming we must first prove our memory  $M$  (Definition 1) has a special property—"optimal substructure". Optimal substructure means one can take a series of observations, and break into to a smaller number of sub-problems or series. If these series/sub-problems have *optimal substructure* they will inherit the relevant optimality present in the original series. That is, if the full series is optimal in some sense all its sub-problems are optimal too. The question for any series is then, "What are the sub-problems that inherit optimality". By proving we can decompose  $M$  optimally—that it has optimal substructure—we also prove that we can grow the series optimally; induction proofs become trivial when the optimal substructure is known.

We prove the optimal substructure of  $M$  of a sequence of state observations. To make these observation we must introduce additional notation. We've previously introduced time  $t$  as an index of observed states, but introduced no mechanism for making state observations.

We formalize the state observation process with two functions, an action policy function  $\pi$  and a transition function  $\delta$ . In informal terms the transition function is an abstract stand in for the world or environment in which animal lives, and the policy represents it's total capacity for value-based decision making.

**Definition 3.** Let  $\pi$  be a policy function that maps a state  $s$  to an action  $a$ ,  $\pi : s \rightarrow a$ , such that  $s \in S, a \in A$ .

**Definition 4.** Let  $\delta$  be a transition function which maps  $(s_t, a_t)$  to a new state  $s_{t+1}$ ,  $\delta : (s_t, a_t) \rightarrow s_{t+1}$ .

By combining policy function  $\pi$  with transition function  $\delta$ , we can generate a sequence of observations, adding these to the memory  $M$ . Given some initial state  $s_t$ , apply  $\pi$  to produce an action  $a_t$ ,  $a_t = \pi(s_t)$ . Given  $a_t$  and  $s_t$  we can apply the transition function, to produce the next observation,

$s_{t+1} = \delta(s_t, a_t)$ . This cycle then repeats, generating a sequence of observations indexed, as we've stated, by  $t$ .

For consistency with standard notation for dynamic programming and the Bellman equation (which comes into play below) we redefine  $E$  in terms of a classic payoff function,  $F(M_t, a_t)$ . We defined  $E$  axiomatically based exclusively on  $M$  and  $d$ . We will use Definitions 3 and 4 to redefine  $E$  in terms the current memory  $M_t$  and the next action, as set by  $\pi$ . This brings us to Definition 5. The axiomatic form of  $E(M_{t+1}, M_t)$  and the payoff form  $F(M_t, a)$  are interchangeable. We switch between there notation as needed.

**Definition 5.** Let  $F(M, a)$  be payout function for  $E$  as defined by Eq. 2.

$$\begin{aligned} F(M_t, a_t) &= E(M_{t+1}, M_t) \\ \text{subject to the constraints} \\ a_t &= \pi(s_t) \\ s_{t+1} &= \delta(s_t, a_t), \\ M_{t+1} &= f(M_t, s_t) \end{aligned} \quad [2]$$

Using Eq 2 we can write down the total information value, as the sum of all payout functions for some policy  $\pi_E$  made over some observation period  $T$ , where  $t = (1, 2, 3, \dots, T)$ . This is value term our dynamic programming solution will maximize.

$$V_{\pi_E} = \sum_{t \in T} F(M_t, a_t) \quad [3]$$

**A proof of optimal substructure.** A first step in solving a dynamic programming problem is isolating the relevant sub-problem, and proving it has an optimal substructure – that the problem can be decomposed into an iteratively optimal series.

**Theorem 1** (Optimal substructure). *Assuming transition function  $\delta$  is deterministic, if  $V_{\pi_E}^*$  is the optimal information value given by  $\pi_E$ , a memory  $M_{t+1}$  has optimal substructure if the the last observation  $s_t$  can be removed from  $M_t$ , by  $M_{t+1} = f^{-1}(M_{t+1}, s_t)$  where the resulting value  $V_{t-1}^* = V_t^* - F(M_t, a_t)$  is also optimal.*

*Proof.* Given a known optimal value  $V^*$  given by  $\pi_E$  we assume for the sake of contradiction there also exists an alternative policy  $\hat{\pi}_E \neq \pi_E$  that gives a memory  $\hat{M}_{t-1} \neq M_{t-1}$  and for which  $\hat{V}_{t-1}^* > V_{t-1}^*$ .

To recover the known optimal memory  $M_t$  we lift  $\hat{M}_{t-1}$  to  $M_t = f(\hat{M}_{t-1}, s_t)$ . This implies  $\hat{V}^* > V^*$  which in turn contradicts the purported original optimality of  $V^*$  and therefore  $\hat{\pi}_E$ .  $\square$

## Information value and search

**An assumption of learning progress.** The study of information geometry models learning as an approach to information equilibrium. As learning progress plateaus, variance decreases, and therefore so does entropy. The rate at which learning plateaus is governed by the Fisher information metric, which is the hessian of the KL-divergence.

Exploration depends on learning progress. We assume that every every state observation is accompanied by change in  $M$ . This change can be arbitrarily small, but must be non-zero.



**Sorting preliminaries.** Our proofs for exploration are really sorting problems. If every state must be visited (or revisited) until learned, then under a greedy policy every state's value must—at one time or another—be the maximum value.

Sorting requires ranking. Ranking requires that we make a definition for both the greater  $>$  and less than inequalities  $<$ . For three real numbers,  $a, b, c \in \mathbb{R}$ .

**Definition 6.**

$$a \leq b \Leftrightarrow \exists c; b = a + c \quad [4]$$

$$a > b \Leftrightarrow (a \neq b) \wedge (b \leq a) \quad [5]$$

**A greedy information policy will explore completely.**

**Definition 7.** Let  $Z$  be set of all visited states, where  $Z_0$  is the empty set  $\{\}$  and  $Z$  is built iteratively over a path  $P$ , such that  $Z = \{s | s \in P \text{ and } s \notin Z\}$ .

**Theorem 2** (State search – completeness and uniqueness). A greedy policy  $\pi$  is the only deterministic policy which ensures all states in  $S$  are visited, such that  $Z = S$ .

*Proof.* Let  $\mathbf{E} = (E_1, E_2, \dots)$  be ranked series of  $E$  values for all states  $S$ , such that  $(E_1 \geq E_2, \geq \dots)$ . To swap any pair of values  $(E_i \geq E_j)$  so  $(E_i \leq E_j)$  by Def. 6  $E_i - c = E_j$ .

Therefore, again by Def. 6,  $\exists \int \delta E(s) \rightarrow -c$ .

Recall:  $\nabla \mathcal{L}_M < 0$

However if we wished to instead swap  $(E_i \leq E_j)$  so  $(E_i \geq E_j)$  by definition  $\nexists c; E_i + c = E_j$ , as  $\nexists \int \delta \rightarrow c$ .

To complete the proof, assume that some policy  $\hat{\pi}_E \neq \pi_E^*$ . By definition policy  $\hat{\pi}_E$  can any action but the maximum leaving  $k - 1$  options. Eventually as  $t \rightarrow T$  the only possible swap is between the max option and the  $k$ th but as we have already proven this is impossible as long as  $\nabla \mathcal{L}_M < 0$ . Therefore, the policy  $\hat{\pi}_E$  will leave at least 1 option unexplored and  $S \neq Z$ .  $\square$

**Theorem 3** (State search – convergence). Assuming a deterministic transition function  $\Lambda$ , a greedy policy  $\pi_E$  will resample  $S$  to convergence as  $t \rightarrow T$ ,  $E_t \rightarrow 0$ .

*Proof.* Recall:  $\nabla \mathcal{L}_M < 0$ .

Each time  $\pi_E^*$  visits a state  $s$ , so  $M \rightarrow M'$ ,  $F(M', a_{t+1}) < F(M, a_t)$

In Theorem 2 we proved only deterministic greedy policy will visit each state in  $S$  over  $T$  trials.

By induction, if  $\pi^* E$  will visit all  $s \in S$  in  $T$  trials, it will revisit them in  $2T$ , therefore as  $T \rightarrow \infty$ ,  $E \rightarrow 0$ .  $\square$

If the transition function  $\Lambda$  is stochastic, the noisy state changes will prevent  $E$  from fully converging to 0. This might be ideal, as it will force continual re-exploration of the world. However if we redefine the converge of  $E$  not to 0 but to some criterion  $\epsilon$ , we can once again ensure convergence in noisy worlds. That is, in the limit of  $t \rightarrow T$ ,  $E_t \rightarrow \epsilon$ , where  $0 \leq \epsilon \ll E_0$  with  $E_0$  denoting the initial value of  $E$ . Too large though, and  $\epsilon$  will interfere with potential optimality of  $\pi_E^*$  (by Theorem. 2).