Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber

IDSIA, Galleria 2, Manno, CH-6928, Switzerland
{yi,tino,juergen}@idsia.ch

Abstract. To maximize its success, an AGI typically needs to explore its initially unknown world. Is there an optimal way of doing so? Here we derive an affirmative answer for a broad class of environments.

1 Introduction

An intelligent agent is sent to explore an unknown environment. Over the course of its mission, the agent makes observations, carries out actions, and incrementally builds up a model of the environment from this interaction. Since the way in which the agent selects actions may greatly affect the efficiency of the exploration, the following question naturally arises:

> *How should the agent choose the actions such that the knowledge about the environment accumulates as quickly as possible?*

In this paper, this question is addressed under a classical framework in which the agent improves its model of the environment through probabilistic inference, and learning progress is measured in terms of Shannon information gain. We show that the agent can, at least in principle, optimally choose actions based on previous experiences, such that the cumulative expected information gain is maximized.

The rest of the paper is organized as follows: Section 2 reviews the basic concepts and establishes the terminology; Section 3 elaborates the principle of optimal Bayesian exploration; Section 4 presents a simple experiment; Related work is briefly reviewed in Section 5; Section 6 concludes the paper.

2 Preliminaries

Suppose that the agent interacts with the environment in discrete time cycles $t = 1, 2, \ldots$. In each cycle, the agent performs an action, a, then receives a sensory input, o. A *history*, h, is either the empty string, $\emptyset$, or a string of the form $a_1 o_1 \cdots a_t o_t$ for some t, and ha and hao refer to the strings resulting from appending a and ao to h, respectively.

2.1 Learning from Sequential Interactions

To facilitate the subsequent discussion under a probabilistic framework, we make the following assumptions:

Assumption I. The models of the environment under consideration are fully described by a random element Θ which *depends solely on the environment.* Moreover, the agent's initial knowledge about Θ is summarized by a prior density $p(\theta)$.

Assumption II. The agent is equipped with a *conditional predictor* $p(o|ha;\theta)$, i.e. the agent is capable of refining its prediction in the light of information about Θ.

Using $p(\theta)$ and $p(o|ha;\theta)$ as building blocks, it is straightforward to formulate learning in terms of probabilistic inference. From Assumption **I**, given the history h, the agent's knowledge about Θ is fully summarized by $p(\theta|h)$. According to Bayes rule, $p(\theta|hao) = \frac{p(\theta|ha)p(o|ha;\theta)}{p(o|ha)}$, with $p(o|ha) = \int p(o|ha,\theta)\, p(\theta|h)\, d\theta$. The term $p(\theta|ha)$ represents the agent's current knowledge about Θ given history h and an additional action a. Since Θ depends solely on the environment, and, importantly, *knowing the action without subsequent observations cannot change the agent's state of knowledge about* Θ, then $p(\theta|ha) = p(\theta|h)$, and thus the knowledge about Θ can be updated using

$$p(\theta|hao) = p(\theta|h) \cdot \frac{p(o|ha;\theta)}{p(o|ha)}. \tag{1}$$

It is worth pointing out that $p(o|ha;\theta)$ is chosen before entering the environment. It is not required that it match the true dynamics of the environment, but the effectiveness of the learning certainly depends on the choices of $p(o|ha;\theta)$. For example, if $\Theta \in \mathbb{R}$, and $p(o|ha;\theta)$ depends on θ only through its sign, then no knowledge other than the sign of Θ can be learned.

2.2 Information Gain as Learning Progress

Let h and h' be two histories such that h is a prefix of h'. The respective posterior distributions of Θ are $p(\theta|h)$ and $p(\theta|h')$. Using h as a reference point, the amount of information gained when the history grows to h' can be measured using the KL divergence between $p(\theta|h)$ and $p(\theta|h')$. This *information gain* from h to h' is defined as

$$g(h'\|h) = KL\left(p(\theta|h')\,\|p(\theta|h)\right) = \int p(\theta|h') \log \frac{p(\theta|h')}{p(\theta|h)} d\theta.$$

As a special case, if $h = \emptyset$, then $g(h') = g(h'\|\emptyset)$ is the *cumulative information gain* with respect to the prior $p(\theta)$. We also write $g(ao\|h)$ for $g(hao\|h)$, which denotes the information gained from an additional action-observation pair.

From an information theoretic point of view, the KL divergence between two distributions p and q represents the additional number of bits required to encode elements sampled from p, using optimal coding strategy designed for q. This can be interpreted as the degree of 'unexpectedness' or 'surprise' caused by observing samples from p when expecting samples from q.

The key property information gain for the treatment below is the following decomposition: Let h be a prefix of h' and h' be a prefix of h'', then

$$\mathbb{E}_{h''|h'} g(h''\|h) = g(h'\|h) + \mathbb{E}_{h''|h'} g(h''\|h'). \tag{2}$$

That is, the information gain is *additive in expectation.*

Having defined the information gain from trajectories ending with observations, one may proceed to define the *expected information gain* of performing action a, before observing the outcome o. Formally, the *expected information gain* of performing a with respect to the current history h is given by $\bar{g}\left(a \| h\right) = \mathbb{E}_{o|ha} g\left(ao \| h\right)$. A simple derivation gives

$$\bar{g}\left(a \| h\right) = \sum_{o} \int p\left(o, \theta | ha\right) \log \frac{p\left(o, \theta | ha\right)}{p\left(\theta | ha\right) p\left(o | ha\right)} d\theta = I\left(O; \Theta | ha\right),$$

which means that $\bar{g}\left(a \| h\right)$ is the mutual information between Θ and the random variable O representing the unknown observation, conditioned on the history h and action a.

3 Optimal Bayesian Exploration

In this section, the general principle of optimal Bayesian exploration in dynamic environments is presented. We first give results obtained by assuming a fixed limited life span for our agent, then discuss a condition required to extend this to infinite time horizons.

3.1 Results for Finite Time Horizon

Suppose that the agent has experienced history h, and is about to choose τ more actions in the future. Let π be a policy mapping the set of histories to the set of actions, such that the agent performs a with probability $\pi\left(a|h\right)$ given h. Define the *curiosity Q-value* $q_{\pi}^{\tau}\left(h, a\right)$ as the expected information gained from the additional τ actions, assuming that the agent performs a in the next step and follows policy π in the remaining $\tau - 1$ steps. Formally, for $\tau = 1$,

$$q_{\pi}^{1}\left(h, a\right) = \mathbb{E}_{o|ha} g\left(ao \| h\right) = \bar{g}\left(a \| h\right),$$

and for $\tau > 1$,

$$\begin{aligned} q_{\pi}^{\tau}\left(h, a\right) &= \mathbb{E}_{o|ha} \mathbb{E}_{a_1|hao} \mathbb{E}_{o_1|haoa_1} \cdots \mathbb{E}_{o_{\tau-1}|h \cdots a_{\tau-1}} g\left(haoa_1o_1 \cdots a_{\tau-1}o_{\tau-1} \| h\right) \\ &= \mathbb{E}_{o|ha} \mathbb{E}_{a_1o_1 \cdots a_{\tau-1}o_{\tau-1}|hao} g\left(haoa_1o_1 \cdots a_{\tau-1}o_{\tau-1} \| h\right). \end{aligned}$$

The curiosity Q-value can be defined recursively. Applying Eq. 2 for $\tau = 2$,

$$\begin{aligned} q_{\pi}^{2}\left(h, a\right) &= \mathbb{E}_{o|ha} \mathbb{E}_{a_1o_1|hao} g\left(haoa_1o_1 \| h\right) \\ &= \mathbb{E}_{o|ha} \left[g\left(ao \| h\right) + \mathbb{E}_{a_1o_1|hao} g\left(a_1o_1 \| hao\right) \right] \\ &= \bar{g}\left(a \| h\right) + \mathbb{E}_{o|ha} \mathbb{E}_{a'|hao} q_{\pi}^{1}\left(hao, a'\right). \end{aligned}$$

And for $\tau > 2$,

$$\begin{aligned} q_{\pi}^{\tau}\left(h, a\right) &= \mathbb{E}_{o|ha} \mathbb{E}_{a_1o_1 \cdots a_{\tau-1}o_{\tau-1}|hao} g\left(haoa_1o_1 \cdots a_{\tau-1}o_{\tau-1} \| h\right) \\ &= \mathbb{E}_{o|ha} \left[g\left(ao \| h\right) + \mathbb{E}_{a_1o_1 \cdots a_{\tau-1}o_{\tau-1}} g\left(haoa_1o_1 \cdots a_{\tau-1}o_{\tau-1} \| hao\right) \right] \\ &= \bar{g}\left(a \| h\right) + \mathbb{E}_{o|ha} \mathbb{E}_{a'|hao} q_{\pi}^{\tau-1}\left(hao, a'\right). \end{aligned} \tag{3}$$

Noting that Eq.3 bears great resemblance to the definition of state-action values ($Q(s,a)$) in reinforcement learning, one can similarly define the *curiosity value* of a particular history as $v_\pi^\tau(h) = \mathbb{E}_{a|h} q_\pi^\tau(h,a)$, analogous to state values ($V(s)$), which can also be iteratively defined as $v_\pi^1(h) = \mathbb{E}_{a|h}\bar{g}(a\|h)$, and

$$v_\pi^\tau(h) = \mathbb{E}_{a|h}\left[\bar{g}(a\|h) + \mathbb{E}_{o|ha} v_\pi^{\tau-1}(hao)\right].$$

The curiosity value $v_\pi^\tau(h)$ is the expected information gain of performing the additional τ steps, assuming that the agent follows policy π. The two notations can be combined to write

$$q_\pi^\tau(h,a) = \bar{g}(a\|h) + \mathbb{E}_{o|ha} v_\pi^{\tau-1}(hao). \tag{4}$$

This equation has an interesting interpretation: since the agent is operating in a dynamic environment, it has to take into account not only the immediate expected information gain of performing the current action, i.e., $\bar{g}(a\|h)$, but also the expected curiosity value of the situation in which the agent ends up due to the action, i.e., $v_\pi^{\tau-1}(hao)$. As a consequence, *the agent needs to choose actions that balance the two factors in order to improve its total expected information gain.*

Now we show that there is a optimal policy π_*, which leads to the maximum cumulative expected information gain given any history h. To obtain the optimal policy, one may work backwards in τ, taking greedy actions with respect to the curiosity Q-values at each time step. Namely, for $\tau = 1$, let

$$q^1(h,a) = \bar{g}(a\|h), \ \pi_*^1(h) = \arg\max_a \bar{g}(a\|h), \text{ and } v^1(h) = \max_a \bar{g}(a\|h),$$

such that $v^1(h) = q^1\left(h, \pi_*^1(h)\right)$, and for $\tau > 1$, let

$$q^\tau(h,a) = \bar{g}(a\|h) + \mathbb{E}_{o|ha}\left[\max_{a'} q^{\tau-1}(a'|hao)\right] = \bar{g}(a\|h) + \mathbb{E}_{o|ha} v^{\tau-1}(hao),$$

with $\pi_*^\tau(h) = \arg\max_a q^\tau(h,a)$ and $v^\tau(h) = \max_a q^\tau(h,a)$. We show that $\pi_*^\tau(h)$ is indeed the optimal policy for any given τ and h in the sense that the curiosity value, when following π_*^τ, is maximized. To see this, take any other strategy π, first notice that

$$v^1(h) = \max_a \bar{g}(a\|h) \geq \mathbb{E}_{a|h}\bar{g}(a\|h) = v_\pi^1(h).$$

Moreover, assuming $v^\tau(h) \geq v_\pi^\tau(h)$,

$$\begin{aligned} v^{\tau+1}(h) &= \max_a\left[\bar{g}(a\|h) + \mathbb{E}_{o|ha} v^\tau(hao)\right] \geq \max_a\left[\bar{g}(a\|h) + \mathbb{E}_{o|ha} v_\pi^\tau(hao)\right] \\ &\geq \mathbb{E}_{a|h}\left[\bar{g}(a\|h) + \mathbb{E}_{o|ha} v_\pi^\tau(hao)\right] = v_\pi^{\tau+1}(h). \end{aligned}$$

Therefore $v^\tau(h) \geq v_\pi^\tau(h)$ holds for arbitrary τ, h, and π. The same can be shown for curiosity Q-values, namely, $q^\tau(h,a) \geq q_\pi^\tau(h,a)$, for all τ, h, a, and π.

Now consider that the agent has a fixed life span T. It can be seen that at time t, the agent has to perform $\pi_*^{T-t}(h_{t-1})$ to maximize the expected information gain in the remaining $T-t$ steps. Here $h_{t-1} = a_1 o_1 \cdots a_{t-1} o_{t-1}$ is the history at time t. However, from Eq.2,

$$\mathbb{E}_{h_T|h_{t-1}} g(h_T) = g(h_{t-1}) + \mathbb{E}_{h_T|h_{t-1}} g(h_T\|h_{t-1}).$$

Note that at time t, $g\left(h_{t-1}\right)$ is a constant, thus *maximizing the cumulative expected information gain in the remaining time steps is equivalent to maximizing the expected information gain of the whole trajectory with respect to the prior.* The result is summarized in the following proposition:

Proposition 1. *Let* $q^{1}\left(h,a\right)=\bar{g}\left(a\|h\right)$, $v^{1}\left(h\right)=\max_{a}q^{1}\left(h,a\right)$, *and*

$$q^{\tau}\left(h,a\right)=\bar{g}\left(a\|h\right)+\mathbb{E}_{o|ha}v^{\tau-1}\left(hao\right),\ v^{\tau}\left(h\right)=\max_{a}q^{\tau}\left(h,a\right),$$

then the policy $\pi_{*}^{\tau}\left(h\right)=\arg\max_{a}q^{\tau}\left(h,a\right)$ *is optimal in the sense that* $v^{\tau}\left(h\right)\geq v_{\pi}^{\tau}\left(h\right)$, $q^{\tau}\left(h,a\right)\geq q_{\pi}^{\tau}\left(h,a\right)$ *for any* π, τ, h *and* a.

In particular, for an agent with fixed life span T, *following* $\pi_{*}^{T-t}\left(h_{t-1}\right)$ *at time* $t=1,\ldots,T$ *is optimal in the sense that the expected cumulative information gain with respect to the prior is maximized.*

The definition of the optimal exploration policy is constructive, which means that it can be readily implemented, provided that the number of actions and possible observations is finite so that the expectation and maximization can be computed exactly. However, the cost of computing such a policy is $O\left(\left(n_{o}n_{a}\right)^{\tau}\right)$, where n_{o} and n_{a} are the number of possible observations and actions, respectively. Since the cost is exponential on τ, planning with large number of look ahead steps is infeasible, and approximation heuristics must be used in practice.

3.2 Non-triviality of the Result

Intuitively, the recursive definition of the curiosity (Q) value is simple, and bears clear resemblance to its counterpart in reinforcement learning. It might be tempting to think that the result is nothing more than solving the finite horizon reinforcement learning problem using $\bar{g}\left(a\|h\right)$ or $g\left(ao\|h\right)$ as the reward signals. However, this is not the case.

First, note that the decomposition Eq.2 is a direct consequence of the formulation of the KL divergence. The decomposition does not necessarily hold if $g\left(h\right)$ is replaced with other types of measures of information gain.

Second, it is worth pointing out that $g\left(ao\|h\right)$ and $\bar{g}\left(a\|h\right)$ behave differently from normal reward signals in the sense that they are *additive only in expectation*, while in the reinforcement learning setup, the reward signals are usually assumed to be additive, i.e., adding reward signals together is always meaningful. Consider a simple problem with only two actions. If $g\left(ao\|h\right)$ is a plain reward function, then $g\left(ao\|h\right)+g\left(a'o'\|hao\right)$ should be meaningful, no matter if a and o is known or not. But this is not the case, since the sum does not have a valid information theoretic interpretation. On the other hand, the sum is meaningful *in expectation.* Namely, when o has *not* been observed, from Eq.2,

$$g\left(ao\|h\right)+\mathbb{E}_{o'|haoa'}g\left(a'o'\|hao\right)=\mathbb{E}_{o'|haoa'}g\left(aoa'o'\|h\right),$$

the sum can be interpreted as the expectation of the information gained from h to $haoa'o'$. This result shows that $g\left(ao\|h\right)$ and $\bar{g}\left(a\|h\right)$ can be treated as additive reward signals only when one is planning ahead.

To emphasize the difference further, note that all immediate information gains $g\left(ao\|h\right)$ are non-negative since they are essentially KL divergence. A natural assumption would be that the information gain $g\left(h\right)$, which is the sum of

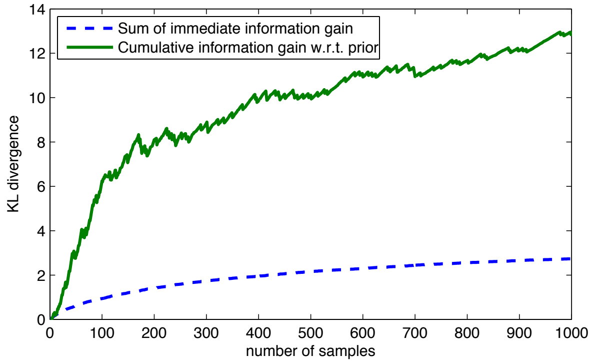

Fig. 1. Illustration of the difference between the sum of one-step information gain and the cumulative information gain with respect to the prior. In this case, 1000 independent samples are generated from a distribution over finite sample space $\{1, 2, 3\}$, with $p\,(x = 1) = 0.1$, $p\,(x = 2) = 0.5$, and $p\,(x = 3) = 0.4$. The task of learning is to recover the mass function from the samples, assuming a Dirichlet prior $Dir\left(\frac{50}{3}, \frac{50}{3}, \frac{50}{3}\right)$. The KL divergence between two Dirichlet distributions are computed according to [5]. It is clear from the graph that the cumulative information gain fluctuates when the number of samples increases, while the sum of the one-step information gain increases monotonically. It also shows that the difference between the two quantities can be large.

all $g\,(ao\|h)$ in expectation, grows monotonically when the length of the history increases. However, this is not the case, see Figure 1 for example. Although $g\,(ao\|h)$ is always non-negative, some of the gain may pull θ closer to its prior density $p\,(\theta)$, resulting in a decrease of KL divergence between $p\,(\theta|h)$ and $p\,(\theta)$. This is never the case if one considers the normal reward signals in reinforcement learning, where the accumulated reward would never decrease if all rewards are non-negative.

3.3 Extending to Infinite Horizon

Having to restrict the maximum life span of the agent is rather inconvenient. It is tempting to define the curiosity Q-value in the infinite time horizon case as the limit of curiosity Q-values with increasing life spans, $T \to \infty$. However, this cannot be achieved without additional technical constraints. For example, consider simple coin tossing. Assuming a $Beta\,(1, 1)$ over the probability of seeing heads, then the expected cumulative information gain for the next T flips is given by

$$v^T\,(h_1) = I\,(\Theta; X_1, \ldots, X_T) \sim \log T.$$

With increasing T, $v^T\,(h_1) \to \infty$. A frequently used approach to simplifying the math is to introduce a discount factor γ ($0 \leq \gamma < 1$), as used in reinforcement learning. Assume that the agent has a maximum τ actions left, but before finishing the τ actions it may be forced to leave the environment with probability $1 - \gamma$ at each time step. In this case, the curiosity Q-value becomes

$q_{\pi}^{\gamma,1}(h,a) = \bar{g}(a\|h)$, and

$$\begin{aligned} q_{\pi}^{\gamma,\tau}(h,a) &= (1-\gamma)\,\bar{g}(a\|h) + \gamma\left[\bar{g}(a\|h) + \mathbb{E}_{o|ha}\mathbb{E}_{a'|hao} q_{\pi}^{\gamma,\tau-1}(hao,a')\right] \\ &= \bar{g}(a\|h) + \gamma \mathbb{E}_{o|ha}\mathbb{E}_{a'|hao} q_{\pi}^{\gamma,\tau-1}(hao,a')\,. \end{aligned}$$

One may also interpret $q_{\pi}^{\gamma,\tau}(h,a)$ as a linear combination of curiosity Q-values without the discount,

$$q_{\pi}^{\gamma,\tau}(h,a) = (1-\gamma)\sum_{t=1}^{\tau}\gamma^{t-1} q_{\pi}^{t}(h,a) + \gamma^{\tau} q_{\pi}^{\tau}(h,a)\,.$$

Note that curiosity Q-values with larger look-ahead steps are weighed exponentially less.

The optimal policy in the discounted case is given by

$$q^{\gamma,1}(h,a) = \bar{g}(a\|h)\,,\ v^{\gamma,1}(h) = \max_{a} q^{\gamma,1}(h,a)\,,$$

and

$$q^{\gamma,\tau}(h,a) = \bar{g}(a\|h) + \gamma\mathbb{E}_{o|ha} v^{\gamma,\tau-1}(hao)\,,\ v^{\gamma,\tau}(h) = \max_{a} q^{\gamma,\tau}(h,a)\,.$$

The optimal actions are given by $\pi_*^{\gamma,\tau}(h) = \arg\max_a q^{\gamma,\tau}(h,a)$. The proof that $\pi_*^{\gamma,\tau}$ is optimal is similar to the one for the finite horizon case (section 3.1) and thus is omitted here.

Adding the discount enables one to define the curiosity Q-value in infinite time horizon in a number of cases. However, it is still possible to construct scenarios where such discount fails. Consider a infinite list of bandits. For bandit n, there are n possible outcomes with Dirichlet prior $Dir\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$. The expected information gain of pulling bandit n for the first time is then given by

$$\log n - \psi(2) + \log\left(1+\frac{1}{n}\right) \sim \log n,$$

with $\psi(\cdot)$ being the digamma function. Assume at time t, only the first $e^{e^{2t}}$ bandits are available, thus the curiosity Q-value in finite time horizon is always finite. However, since the largest expected information gain grows at speed e^{t^2}, for any given $\gamma > 0$, $q^{\gamma,\tau}$ goes to infinity with increasing τ. This example gives the intuition that to make the curiosity Q-value meaningful, the 'total information content' of the environment (or its growing speed) must be bounded.

The following technical Lemma gives a sufficient condition for when such extension is meaningful.

Lemma 1. *We have*

$$0 \le q^{\gamma,\tau+1}(h,a) - q^{\gamma,\tau}(h,a) \le \gamma^{\tau}\mathbb{E}_{o|ha}\max_{a_1}\mathbb{E}_{o_1|haoa_1}\cdots\max_{a_\tau}\bar{g}(a_\tau\|h\cdots o_{\tau-1})\,.$$

Proof. Expand $q^{\gamma,\tau}$ and $q^{\gamma,\tau+1}$, and note that $|\max X - \max Y| \leq \max |X - Y|$, then

$$
\begin{aligned}
&q_\pi^{\gamma,\tau+1}(h,a) - q_\pi^{\gamma,\tau}(h,a) \\
&= \mathbb{E}_{o|ha} \max_{a_1} \mathbb{E}_{o_1|haoa_1} \cdots \max_{a_\tau} \left[\bar{g}(a\|h) + \gamma \bar{g}(a_1\|hao) + \cdots + \gamma^\tau \bar{g}(a_\tau\|h\cdots o_{\tau-1})\right] \\
&- \mathbb{E}_{o|ha} \max_{a_1} \mathbb{E}_{o_1|haoa_1} \cdots \max_{a_{\tau-1}} \left[\bar{g}(a\|h) + \gamma \bar{g}(a_1\|hao) + \cdots + \gamma^{\tau-1} \bar{g}(a_{\tau-1}\|h\cdots o_{\tau-2})\right] \\
&\leq \mathbb{E}_{o|ha} \max_{a_1} \{ \mathbb{E}_{o_1|haoa_1} \cdots \max_{a_\tau} \left[\bar{g}(a\|h) + \gamma \bar{g}(a_1\|hao) + \cdots + \gamma^\tau \bar{g}(a_\tau\|h\cdots o_{\tau-1})\right] \\
&- \mathbb{E}_{o_1|haoa_1} \cdots \max_{a_{\tau-1}} \left[\bar{g}(a\|h) + \gamma \bar{g}(a_1\|hao) + \cdots + \gamma^{\tau-1} \bar{g}(a_{\tau-1}\|h\cdots o_{\tau-2})\right] \} \\
&\leq \cdots \\
&\leq \gamma^\tau \mathbb{E}_{o|ha} \max_{a_1} \mathbb{E}_{o_1|haoa_1} \cdots \max_{a_\tau} \bar{g}(a_\tau\|h\cdots o_{\tau-1}) .
\end{aligned}
$$

It can be seen that if $\mathbb{E}_{oa_1\cdots o_{\tau-1}a_\tau|ha} \bar{g}(a_\tau\|h\cdots o_{\tau-1})$ grows sub-exponentially, then $q_\pi^{\gamma,\tau}$ is a Cauchy sequence, and it makes sense to define the curiosity Q-value for infinite time horizon.

4 Experiment

The idea presented in the previous section is illustrated through a simple experiment. The environment is an MDP consisting of two groups of densely connected states (cliques) linked by a long corridor. The agent has two actions allowing it to move along the corridor deterministically, whereas the transition probabilities inside each clique are randomly generated. The agent assumes Dirichlet priors over all transition probabilities, and the goal is to learn the transition model of the MDP. In the experiment, each clique consists of 5 states, (states 1 to 5 and states 56 to 60), and the corridor is of length 50 (states 6 to 55). The prior over each transition probability is $Dir\left(\frac{1}{60}, \ldots, \frac{1}{60}\right)$.

We compare four different algorithms: i) random exploration, where the agent selects each of the two actions with equal probability at each time step; ii) Q-learning with the immediate information gain $g(ao\|h)$ as the reward; iii) greedy exploration, where the agent chooses at each time step the action maximizing $\bar{g}(a\|h)$; and iv) a dynamic-programming (DP) approximation of the optimal Bayesian exploration, where at each time step the agent follows a policy which is computed using policy iteration, assuming that the dynamics of the MDP is given by the current posterior, and the reward is the expected information gain $\bar{g}(a\|h)$. The detail of this algorithm is described in [11].

Fig.2 shows the typical behavior of the four algorithms. The upper four plots show how the agent moves in the MDP starting from one clique. Both greedy exploration and DP move back and forth between the two cliques. Random exploration has difficulty moving between the two cliques due to the random walk behavior in the corridor. Q-learning exploration, however, gets stuck in the initial clique. The reason for is that since the jump on the corridor is deterministic, the information gain decreases to virtually zero after only several attempts, therefore the Q-value of jumping into the corridor becomes much lower than the

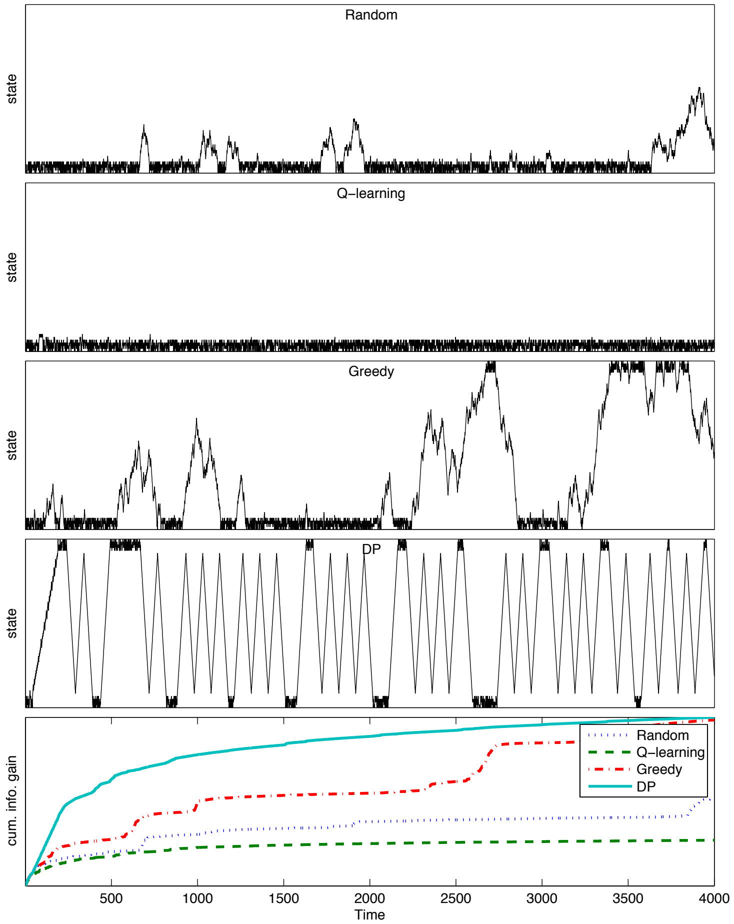

Fig. 2. The exploration process of a typical run of 4000 steps. The upper four plots shows the position of the agent between state 1 (the lowest) and 60 (the highest). The states at the top and the bottom correspond to the two cliques, and the states in the middle correspond to the corridor. The lowest plot is the cumulative information gain with respect to the prior.

Q-value of jumping inside the clique. The bottom plot shows how the cumulative information gain grows over time, and how the DP approximation clearly outperforms the other algorithms, particularly in the early phase of exploration.

5 Related Work

The idea of actively selecting queries to accelerate learning process has a long history [1, 2, 7], and has received a lot of attention in recent decades, primarily in the context of active learning [8] and artificial curiosity [6]. In particular, measuring learning progress using KL divergence dates back to the 50's [2, 4]. In

1995 this was combined with reinforcement learning, with the goal of optimizing future expected information gain [10]. Others renamed this Bayesian surprise [3].

Our work differs from most previous work in two main points: First, like in [10], we consider the problem of exploring a dynamic environment, where actions change the environmental state, while most work on active learning and Bayesian experiment design focuses on queries that do not affect the environment [8]. Second, our result is theoretically sound and directly derived from first principles, in contrast to the more heuristic application [10] of traditional reinforcement learning to maximize the expected information gain. In particular, we pointed out a previously neglected subtlety of using KL divergence as learning progress.

Conceptually, however, this work is closely connected to artificial curiosity and intrinsically motivated reinforcement learning [6,7,9] for agents that actively explore the environment without an external reward signal. In fact, the very definition of the curiosity (Q) value permits a firm connection between pure exploration and reinforcement learning.

6 Conclusion

We have presented the principle of optimal Bayesian exploration in dynamic environments, centered around the concept of the curiosity (Q) value. Our work provides a theoretically sound foundation for designing more effective exploration strategies. Future work will concentrate on studying the theoretical properties of various approximation strategies inspired by this principle.

7 Acknowledgement

This research was funded in part by Swiss National Science Foundation grant 200020-122124/1, and the EU IM-CLeVeR project (#231722).

References

1. Chaloner, K., Verdinelli, I.: Bayesian experimental design: A review. Statistical Science 10, 273–304 (1995)
2. Fedorov, V.V.: Theory of optimal experiments. Academic Press (1972)
3. Itti, L., Baldi, P.F.: Bayesian surprise attracts human attention. In: NIPS'05. pp. 547–554 (2006)
4. Lindley, D.V.: On a measure of the information provided by an experiment. Annals of Mathematical Statistics 27(4), 986–1005 (1956)
5. Penny, W.: Kullback-liebler divergences of normal, gamma, dirichlet and wishart densities. Tech. rep., Wellcome Department of Cognitive Neurology, University College London (2001)
6. Schmidhuber, J.: Curious model-building control systems. In: IJCNN'91. vol. 2, pp. 1458–1463 (1991)
7. Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation (1990-2010). Autonomous Mental Development, IEEE Trans. on Autonomous Mental Development 2(3), 230–247 (9 2010)
8. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin Madison (2010)
9. Singh, S., Barto, A., Chentanez, N.: Intrinsically motivated reinforcement learning. In: NIPS'04 (2004)
10. Storck, J., Hochreiter, S., Schmidhuber, J.: Reinforcement driven information acquisition in non-deterministic environments. In: ICANN'95 (1995)
11. Sun, Y., Gomez, F.J., Schmidhuber, J.: Planning to be surprised: Optimal bayesian exploration in dynamic environments (2011), http://arxiv.org/abs/1103.5708