LEARNING TO EXPLORE WITH PLEASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Exploration is a long-standing challenge in sequential decision problem in machine learning. This paper investigates the adoption of two theories of optimal stimulation level - "the pacer principle" and the Wundt curve - from psychology to improve the exploration challenges. We propose a method called exploration with pleasure (EP) which is formulated based on the notion of pleasure as defined in accordance with the above two theories. EP is able to identify the region of stimulations that will trigger pleasure to the learning agent during exploration and consequently improve on the learning process. The effectiveness of EP is studied in two machine learning settings: curiosity-driven reinforcement learning (RL) and Bayesian optimisation (BO). Experiments in purely curiosity-driven RL show that by using EP to generate intrinsic rewards, it can yield faster learning. Experiments in BO demonstrate that by using EP to specify the exploration parameters in two acquisition functions - Probability of Improvement and Expected Improvement - it can achieve faster convergence and better function values.

1 INTRODUCTION

In psychology, theories of curiosity (Franken, 2006; Kidd & Hayden, 2015) seek to explain the causes of organism's exploratory behaviours. One research strand of theories of curiosity explains that organism's exploratory behaviours are motivated by stimulus's properties such as novelty, surprisingness. Theories of curiosity have been widely adopted by researchers to guide exploration in learning systems, for example in intrinsically-motivated reinforcement learning (RL) (Still & Precup, 2012; Barto, 2013; Stadie et al., 2015; Mohamed & Jimenez Rezende, 2015; Houthooft et al., 2016; Bellemare et al., 2016; Pathak et al., 2017; Achiam & Sastry, 2017; Burda et al., 2019a). These approaches are known as curiosity-driven exploration. Theories of curiosity in psychology are commonly associated with theories of optimal stimulation level (Dember & Earl, 1957; Berlyne, 1966; 1978; Zuckerman, 2016) which are equivalently important in explaining organism's exploratory behaviour. Theories of optimal stimulation level state that organism achieves maximum enjoyment when the stimulation is at intermediate level, i.e. not too low that causes boredom, and not too high that causes anxiety. These theories have been central to the study of infant and child curiosity and learning (Dember & Earl, 1957; Berlyne, 1978; Kinney & Kagan, 1976; Kidd et al., 2012). Nevertheless, theories of optimal stimulation level are less studied by machine learning researchers.

This paper investigates the adoption of two theories of optimal stimulation level from psychology - "the pacer principle" (Dember & Earl, 1957) and the Wundt curve (Berlyne, 1978) - and studies in what ways the adoption might help to enhance the exploration strategies in machine learning, specifically in sequential decision problem. We present a method, known as exploration with pleasure (EP), which is formulated in accordance with the "pacer principle" and the Wundt curve. EP is able to identify the region of stimulations that will invoke pleasure to the learning agent and hence encourage exploration and learning. The effectiveness of EP is demonstrated in two machine learning settings, i.e curiosity-driven reinforcement learning (RL) and Bayesian optimisation (BO).

2 BACKGROUND

Exploratory behaviour in psychology literature. Curiosity is a kind of intrinsic motivation (Silvia, 2012) that promotes exploratory and manipulatory behaviours. One strand of research regarding the causes of curiosity explains that curiosity in organsim is aroused by external stimuli in the environment (Franken, 2006; Zuckerman, 2016). Berlyne's theory of curiosity (Berlyne, 1960; 1966;

1971; 1973; 1978) suggests that there are three classes of stimulus properties that will arouse curiosity: psychophysiological properties (e.g. intensity, colour, and pitch), ecological properties, and collative or structural properties (e.g. novelty-familiarity, simplicity-complicity, clarity-obscurity, and expectedness-surprisingness). A stimulus property attracts an organism's preference and brings intangible or intrinsic rewards, such as pleasantness, to the organism. Such intangible rewards are collectively known as hedonic value. In parallel with theories of curiosity, theories of optimal stimulation level describe that there are certain optimal stimulation levels that will provide the organism with maximum hedonic values (Zuckerman, 2016). Two influential theories are "the pacer principle" proposed by Dember and Earl (Dember & Earl, 1957), and the Wundt curve proposed by Berlyne's (Berlyne, 1960; 1971). "The pacer principle" (Dember & Earl, 1957) states that when organisms get used to a certain level of arousal potential (called an adaptation level), they will become bored and will seek to explore stimuli with slightly higher level of stimulation than the adaptation level. In other words, stimuli with increasing higher stimulation will invoke higher hedonic value. The Wundt curve is an inverted U-shaped curve which explains that hedonic value increases with the increase of stimulation up to a maximum point after which the hedonic value decreases with any further increase of stimulation. The curve shows that high hedonic values are triggered by some intermediate level of stimulation. In this paper, we adopt "the pacer principle" and the Wundt curve to construct a method known as exploration with pleasure (EP).

Exploration in curiosity-driven reinforcement learning (RL). Curiosity-driven RL is a type of intrinsically-motivated reinforcement learning (Barto, 2013) which adopts psychological theories of curiosity. The key idea of curiosity-driven exploration in RL is to encourage the agent to explore states (i.e. the stimuli) that exhibit arousing properties such as novelty, surprisingness etc. Upon visiting these states, the agent is awarded some intrinsically-generated rewards known as intrinsic rewards or exploration bonuses. In the RL literature, the most studied stimulus properties are novelty and surprisingness. The main challenge is to quantify these properties within a state. Researchers have proposed a range of methods to quantify the novelty of a state as: (1) visitation count (Bellemare et al., 2016; Ostrovski et al., 2017; Tang et al., 2017), or (2) distance between a state and other states (Fu et al., 2017; Savinov et al., 2019; Kim et al., 2019). Methods to quantify the surprisingness of a state include the prediction error methods (Stadie et al., 2015; Pathak et al., 2017; Burda et al., 2019a;b), and the information gain methods (Still & Precup, 2012; Houthooft et al., 2016; Achiam & Sastry, 2017). Existing research on curiosity-driven exploration focus extensively on the methods to quantify a state's properties (i.e. novelty, surprisingness) that will arouse curiosity and compute intrinsic reward as proportional to the quantification. These studies aim to seek as much novelty and surprisingness as possible. Less attention has been paid to investigation concerning the strength or "impact" of stimulation invoked by the state properties. This aspect is studied in this paper with the adoption of the theories of optimal stimulation level. We demonstrate the use of EP to generate intrinsic rewards in section 4.

Exploration in Bayesian optimisation (BO). BO (Brochu et al., 2010; Shahriari et al., 2016) is a probabilistic and sample-efficient approach to global optimisation of black-box objective functions. The goal of global optimisation is to find the maxima or the minima point of the objective function in a data efficient way, using a low number of function evaluations. Two main components in BO are the surrogate model and the acquisition function. A surrogate model provides a probabilistic belief for the objective function conditioned on a sequence of observed data. The role of the acquisition function is to guide the search for the optimum. At each iteration of optimisation, the acquisition function leverages the uncertainty in the posterior belief provided by the surrogate model to select the next query point to evaluate. Two common types of acquisition functions are Probability of Improvement (PI) (Kushner, 1964) and Expected Improvement (EI) (Močkus, 1975). These acquisition functions use some exploration parameters to control the exploration during the selection of next query points in the optimisation procedure. The exploration parameters are commonly set heuristically and left to the user. This paper proposes the use of EP to specify the exploration parameters, as presented in section 5.

3 Exploration with Pleasure

This section presents a method of exploration with pleasure (EP) in accordance with "the pacer principle" (Dember & Earl, 1957) and the Wundt curve (Berlyne, 1966; 1971) from psychology.

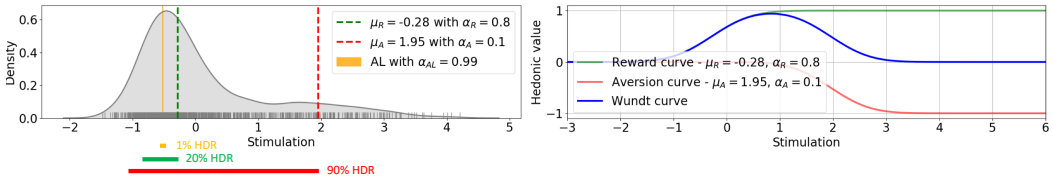

Figure 1: (Left) An example of density plot of a history of stimulations. (Right) The reward, aversion and Wundt curves as identified by using EP.

The goal of EP is to determine the range of stimulations that will invoke pleasure and to compute the amount of pleasure to be enjoyed by a learning agent upon visiting a state and experiencing the stimulation properties exhibited by the state.

Let s be a stimulus which exhibits a range of stimulation properties $\{\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_n\}$. Each stimulation property u_i is defined as a random variable. In psychology, hedonic value of stimuli refers to the pleasantness invoked by the stimuli properties (Laane, 2011). Hedonic value and pleasure are used interchangeably in this paper. Let $h(u)$ be the hedonic value invoked by a stimulation property u_i of a stimulus s. In a sequential decision process, an agent encounters a sequence of states when interacting with the environment. A state corresponds to a stimulus. The stimulation properties of a state determine the hedonic value of the state which motivates the exploratory behaviour of the agent. In other words, the pleasure experienced by the agent upon visiting a state is a function of the state's stimulation properties. We consider a single stimulation property u in this paper. Algorithm 1 shows the steps of computing the hedonic value of a new stimulation conditioned on a history of past stimulations. Each step is explained below.

Algorithm 1 Exploration with pleasure (EP)

Input: stimulations $u_{t-K:t-1}$
Estimate probability density function $\phi_{u_{t-K:t-1}}$
Determine adaptation level AL
Compute reward curve Φ_R
Compute aversion curve Φ_A
Compute hedonic value with Wundt curve: $h(u_t) = W(u_t) = \Phi_R(u_t) - \Phi_A(u_t)$

Definition 3.1. Stimulation density. A stimulation property is defined as a random variable $\mathrm{u} \in \mathbb{R}^+$. A probability density function ϕ_{u}, describes the density of the stimulation variable u. A stimulation u with high density indicates high occurrences of that stimulation. u is analysed by using method of high density regions (HDR), which is a method of selecting a region of the sample space covering a specified probability (Hyndman, 1996). The $100(1-\alpha)\%$ HDR is the subset $R(\phi_\alpha)$ of the sample space of u such that: $R(\phi_\alpha) = \{u : \phi(u) \geq \phi_\alpha\}$, where ϕ_α is the largest constant such that $p(\mathrm{u} \in R(\phi_\alpha)) \geq 1-\alpha$. Figure 1 (left) shows the particular values of $\phi_{0.99}$, $\phi_{0.8}$ and $\phi_{0.1}$, which was used to construct the 1% HDR, 20% HDR and 90% HDR.

The pacer principle (Dember & Earl, 1957) states that when organisms get used to a certain level of stimulus property (called an adaptation level), they will become bored and will seek to explore stimuli with slightly higher level of stimulation than the adaptation level. In other words, stimuli with increasing higher stimulation will invoke higher hedonic value. Formally, the adaptation level (AL) is defined by Definition 3.2 below.

Definition 3.2. Adaptation level. Let $u_{t-K:t-1}$ be the set of stimulations invoked by a sequence of states visited by an agent from time $t-K$ to time $t-1$. Assuming an unimodal distribution, an adaptation level AL is approximately equivalent to the 1% HDR (i.e. $\phi_{0.99}$) of the probability density function $\phi_{u_{t-K:t-1}}$, with $\alpha_{AL} = 0.99$:

$$AL \approx \{u : \phi_{u_{t-K:t-1}}(u) \geq \phi_{\alpha_{AL}}\}$$

The 1% HDR is the region of stimulations that an agent has had high familiarity and hence greater boredom. The region of stimulations that is slightly higher than the 1% HDR is expected to trigger pleasure to the agent.

Berlyne's Wundt curve is based on the psychology studies (Milner, 1991) which have recognised that the reward and aversion systems of the brain assign a motivational valence to any stimuli by determining whether they are rewarding and should be approached or are aversive and should be avoided. The reward system, when activated, assigns positive hedonic values. The aversion system, when activated, assigns negative hedonic values. The arousing properties of a stimulus will determine how intensely it will activate either the reward system or the aversion system. The activation of the aversion system will inhibit the reward system. According to Berlyne's ideas (Berlyne, 1966; 1971), the activation of the reward and aversion systems can be modelled as two cumulative distribution functions, as shown in Figure 1 (right). The reward curve maps stimulus property to positive hedonic value. The aversion curve works in opposite direction, mapping stimulus property to negative hedonic value, and it takes more stimulation to activate. The algebraic sum of the reward curve and aversion curve produces an inverted U-shaped Wundt curve which specifies the region of stimulations that will invoke pleasure to the agent. Formally, these three curves are given by Definition 3.3, 3.4 and 3.5 below.

Definition 3.3. Reward curve. The reward curve determines the region of stimulation that is rewarding to the agent. The reward curve is represented by a normal cumulative density function Φ_R with mean μ_R and standard deviation σ_R. The parameters μ_R and σ_R are formalised as follows:

$$\mu_R = \text{sup}\ \{u : \phi_{u_{t-K:t-1}}(u) \geq \phi_{\alpha_R}\} \tag{1}$$

$$\sigma_R^2 = \text{Var}\ (\text{u}|\ \mu_R < \text{u} < \mu_A) \tag{2}$$

$$\text{s.t.} \quad AL < \mu_R < \mu_A \ \ , \ \ \alpha_A < \alpha_R < \alpha_{AL} \tag{3}$$

The mean of the reward curve μ_R is set to the supremum of the $100(1-\alpha_R)\%$ HDR of $\phi_{u_{t-K:t-1}}$. The value of μ_R is conditioned on $AL < \mu_R < \mu_A$ and the value of α_R is conditioned on $\alpha_A < \alpha_R < \alpha_{AL}$, where α_{AL} is the HDR of the adaptation level (see Definition 3.2), μ_A is the mean of the aversion curve and α_A is the HDR of the aversion curve as defined by Definition 3.4 below. Both of the conditions insist that the region of rewarding stimulations is slightly higher than the adaptation level and lower than the aversive region. The variance of the reward curve σ_R^2 is computed as the variance of the truncated distribution between μ_R and μ_A.

Definition 3.4. Aversion curve. The aversion curve specifies the region of stimulations that will invoke negative hedonic value to the agent. The aversion curve is represented by a normal cumulative density function Φ_A with mean μ_R and standard deviation σ_R formalised as follows:

$$\mu_A = \text{sup}\ \ \{u : \phi_{u_{t-K:t-1}}(u) \geq \phi_{\alpha_A}\} \tag{4}$$

$$\sigma_A^2 = \text{Var}\ (\text{u}|\ \mu_R < \text{u} < \mu_A) \tag{5}$$

$$\text{s.t.} \quad AL < \mu_R < \mu_A \ \ , \ \ \alpha_A < \alpha_R < \alpha_{AL} \tag{6}$$

The mean of the aversion curve μ_A is set to the supremum of the $100(1-\alpha_A)\%$ HDR. The curve is conditioned on $AL < \mu_R < \mu_A$ and $\alpha_A < \alpha_R < \alpha_{AL}$, which means that the aversion curve covers the region of stimulations that are higher than the adaptation level and the reward curve.

Definition 3.5. The Wundt curve. The Wundt curve W is the algebraic sum of the reward curve Φ_R and aversion curve Φ_A. Conditioned on the stimulations $u_{t-K:t-1}$ that have been experienced by an agent, the hedonic value or pleasure that will be triggered by a stimulation at time t, u_t, is computed as:

$$h(u_t) = W(u_t) = \Phi_R(u_t;\ \mu_R, \sigma_R) - \Phi_A(u_t;\ \mu_A, \sigma_A) \tag{7}$$

Figure 1 illustrates the definitions of the adaptation level, the reward curve, the aversion curve and the Wundt curve. From the figure it is clearly seen that α_R determines the distance of the Wundt curve from the adaptation level; the smaller the α_R, the larger the μ_R and hence the further the Wundt curve is distanced from the adaptation level AL. In other words, smaller value of α_R means that higher stimulations are required to invoke pleasure. α_A determines the width of the Wundt curve; the smaller the α_A, the larger the μ_A and hence the wider the Wundt curve is, i.e. the range of stimulations that will invoke pleasure is wider.

Algorithm 2 Computing intrinsic reward with EP

procedure EP($u_{t-K:t-1}$, u_t)
Input: $\alpha_R \leftarrow$ reward curve HDR, $\alpha_A \leftarrow$ aversion curve HDR
Estimate probability density of $u_{t-K:t-1}$: $\phi_{u_{t-K:t-1}}$
Identify adaptation level AL at 1% HDR of $\phi_{u_{t-K:t-1}}$ (**Definition 3.2**)
With ϕ_U, compute reward curve Φ_R parameters: μ_R, σ_R (**Definition 3.3**)
With ϕ_U, compute aversion curve Φ_A parameters: μ_A, σ_A (**Definition 3.4**)
Compute hedonic value $h_t(u_t) = W(u_t) = \Phi_R(u_t) - \Phi_A(u_t)$ (**Definition 3.5**)
return $r_t^i = h_t$

Discussion. The use of Wundt curve to guide an intelligent agent has been reported by Saunders (2002) and Merrick (2013). In their studies, the parameters of the curves were pre-determined and remained static over the course of learning. We distinguish EP from their work in two aspects: (1) EP combines the adoption of "the pacer principle" and the Wundt curve; (2) the parameters of the reward and aversion curves in EP are configured adaptively over the course of learning. The next two sections present the applications of EP in curiosity-driven RL and BO.

4 Application in Curiosity-Driven Reinforcement Learning

4.1 Preliminaries

Consider a purely curiosity-driven RL setting with infinite horizon. At time step t, the agent interacts with the environment by visiting a state s_t, executing an action a_t sampled from its current policy π and moving into next state s_{t+1}. Extrinsic reward r^e is not available, learning is to be driven exclusively by intrinsic reward r^i. Consider surprisingness as the single type of stimulation property u exhibited by a state. The surprisingness of state s_{t+1}, denoted by u_t, is quantified by using the prediction-error-based method. The prediction-error-based method is used in this paper because it has been shown to perform well in purely curiosity-driven RL experiments by Burda et al. (2019a). By using the same method as Burda et al. (2019a), u_t is quantified as the prediction error for a problem related to the agent's transitions. A network is used to encode a state s_t into a feature vector $v(s_t)$. A forward dynamics network parameterised by β, $D_\beta : S \times A \rightarrow S$, is used to predict the encoded next state $\hat{v}(s_{t+1})$ given the current encoded state $v(s_t)$ and action a_t. Surprisingness u_t is computed as the prediction error of the forward model, i.e. the error between the predicted next state $\hat{v}(s_{t+1})$ and the ground truth next state $v(s_{t+1})$, i.e. $u_t = \|D_\beta(v(s_t), a_t) - v(s_{t+1})\|^2$. The algorithm of the purely curiosity-driven RL can be found in Alg. 5 in App. A.1.

4.2 Method

Algorithm 2 show the steps of using EP to generate intrinsic reward r_t^i conditioned on a range of previous surprisingness $u_{t-K:t-1}$. A double-ended buffer U is used to store the last K surprisingness that have been experienced by the agent in the last N_u rollouts of length J, where $K = N_u \times J$. At time t, upon encountering surprisingness u_t, the hedonic value invoked by u_t is computed by using EP. EP relates u_t to hedonic value, i.e. the pleasure to be enjoyed by the agent, by analysing the history of surprisingness contained in buffer U with respect to probability density. Let $\phi_{u_{t-K:t-1}}$ be the probability density of the history of surprisingness in buffer U. Kernel density estimation (KDE) is used to estimate $\phi_{u_{t-K:t-1}}$, which is then analysed with high density region (HDR) method to specify the adaptation level (AL), the parameters of the reward curve Φ_R and aversion curve Φ_A in accordance with Definitions 3.2, 3.4, 3.5. The hedonic value h_t invoked by u_t is computed by the Wundt curve as defined in Definition 3.6. Hedonic value h_t is used as intrinsic reward, i.e. $r_t^i = h_t$.

4.3 Experiments

We evaluated the effectiveness of the EP agent on the Arcade Learning Environment (ALE) (Bellemare et al., 2013). We chose a set of 4 Atari games: Pong, SpaceInvaders, Riverraid and Asterix. These games were chosen because the surprisingness experienced by the agent in these games exhibit four distinguished patterns of density, as described in App. A.2. Our experiments were to

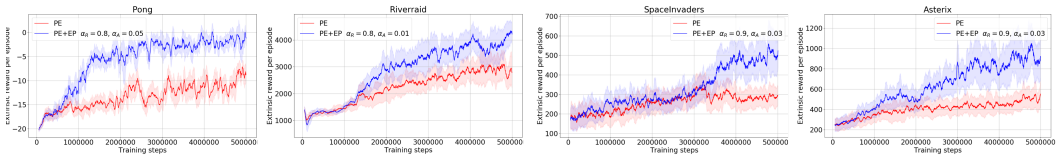

Figure 2: Comparing PE and PE+EP on Pong, Riverraid, SpaceInvaders and Asterix.

Table 1: Results against baseline on Pong, Riverraid, SpaceInvaders and Asterix.

Method	**Pong**	**Riverraid**	**SpaceInvaders**	**Asterix**
PE	-8.8±1.8	2791.4±562.8	294.8±48.7	551.3±124.3
PE+EP	**-1.1±1.6**	**4215.3±383.5**	**523.1±82.8**	**912.9±210.9**

investigate the ability of EP in handling these different types of surprisingness density patterns. All experiments used the Proximal Policy Optimisation (PPO) RL algorithm (Schulman et al., 2017) and the prediction-error-based surprise quantification method. We adopted the same core architecture as that used by Burda et al. (2019a). The hyperparameters of the PPO agent and the surprise quantification method are given in App. A.3. All learning curves are the average of three runs with different seeds over 5 millions time steps. The performance of the agent was measured by the extrinsic rewards achieved by the agent. It is important to note that the extrinsic reward is only used for evaluation, not for training; training is driven exclusively by intrinsic reward.

We conducted experiments to compare prediction-error surprise quantification methods with and without EP. PE denotes method without EP; PE+EP denotes methods with EP. Hyperparameters used in EP are: $\alpha_R \in \{0.9, 0.8\}$, $\alpha_A \in \{0.05, 0.07, 0.01\}$, $K = 512$. These hyperparameters are chosen following a coarse grid search over a range of values, as detailed in App. A.4. Figure 2 illustrates the learning curves and Table 1 summarises the results. It is clearly seen that PE+EP outperformed PE across all 4 games. One significant observation is found in the Pong and Riverraid experiments. In contrast to Burda et al. (2019a) who used 128 parallel environments with 200 million frames, PE+EP was able to achieve similar results by using only 32 parallel environments with 5 million time steps, i.e. 20 million frames, which are 10 times less than Burda et al. (2019a). We conclude two points from our results. First, hedonic value is a better signal to be used as intrinsic reward because EP computes the hedonic value by analysing the impact of surrprisingness. Second, our results enhance the findings of Burda et al. (2019a), who highlighted that there is a high degree of alignment between the intrinsic curiosity objective and the hand-designed extrinsic rewards of many game environments; our results indicate that intrinsic curiosity objective governed by EP is better aligned with the game-designers' extrinsic rewards in the four games we evaluated. Nevertheless, further experiments are required in order to confirm this finding on more Atari games.

5 Application in Bayesian Optimisation

5.1 Preliminaries

Algorithm 3 Bayesian optimisation (BO)

Input: f - unknown objective function, $\mathbb{X}$ - input domain
Input: a_t - acquisition function, T - fixed query budget
Initialise: $D_0 \leftarrow (x_0, y_0)$
for $t = 1, 2, \ldots, T$ **do**
  Fit model M to current data $D_{0:t-1}$
  Select query point: $\boldsymbol{x}_t \leftarrow argmax_{\mathbb{X}}\, a_t(\boldsymbol{x}|D_{0:t-1})$
  Evaluate objective function to obtain $y_t \leftarrow f(\boldsymbol{x}_t)$
  Augment the data $D_{0:t} \leftarrow \{D_{0:t-1}, (\boldsymbol{x}_t, y_t)\}$
return $\boldsymbol{x}^*_T$

Algorithm 3 shows the steps of the BO strategy. Consider a minimisation problem which is to find the minimum point $\boldsymbol{x}^* \in \mathbb{X}$ of an objective function $f : \mathbb{X} \to \mathbb{Y}$ by solving $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{X}} f(\boldsymbol{x})$. The goal is to find the minima point with a low number of function evaluations. $D = \{(\boldsymbol{x}_i), y_i\}$ denotes the available data regarding f, assuming noiseless observation $y = f(\boldsymbol{x})$. BO constructs a prior belief of the f with a surrogate model M. First evaluate the objective function f with a few initial points and fit the surrogate model with the initial observations y via Bayesian posterior updating. At each iteration t, an acquisition function a is used to score a set of sampled query points and the query point with highest score is chosen for the next function evaluation. The new observation is used to update the surrogate model. The selection of next query point and the update of surrogate model are repeated iteratively until a stopping criteria is met. Gaussian Procecss (GP) (Rasmussen & Williams, 2006) is the most widely used surrogate model in BO. In this paper, GP is used exclusively. Two commonly used acquisition functions in BO are the probability of improvement (PI) (Kushner, 1964) and the expected improvement (EI) (Močkus, 1975) methods. Both PI and EI are heuristic approaches. PI and EI compute the score of a query point as follows:

$$a_t^{PI}(\boldsymbol{x}) = p(f(\boldsymbol{x}) < \tau) = \Phi\Big(\frac{\tau - \mu_{t-1}(\boldsymbol{x})}{\sigma_{t-1}(\boldsymbol{x})}\Big) \tag{8}$$

$$a_t^{EI}(\boldsymbol{x}) = (\tau - \mu_{t-1}(\boldsymbol{x}))\ \Phi(\frac{\tau - \mu_{t-1}(\boldsymbol{x})}{\sigma_{t-1}(\boldsymbol{x})}) + \sigma_{t-1}(\boldsymbol{x})\ \phi(\frac{\tau - \mu_{t-1}(\boldsymbol{x})}{\sigma_{t-1}(\boldsymbol{x})}) \tag{9}$$

where $\Phi(z)$ is the standard Gaussian cumulative density function, $\phi(z)$ is the standard Gaussian probability density function, and τ is a threshold value. Both PI and EI have different units of measure: PI measures the probability that a query point $\boldsymbol{x}$ improves upon target τ; EI computes the magnitude of improvement that a query point $\boldsymbol{x}$ is likely to make upon target τ. In practice, a common choice is setting τ to the best observed value, i.e. $\tau = y^{min} := \min_{i=1:t-1}\ y_i$, or the current best incumbent, i.e. $\tau = \mu_t^{min} := \min_{i=1:t-1}\ \mu_{t-1}(\boldsymbol{x}_i)$. To encourage more exploration, a small exploration parameter ξ is commonly added such that $\tau = y^{min} - \xi$, $\tau = \mu_{t-1}^{min} - \xi$. The value of ξ is left to the user.

5.2 METHOD

This section presents an approach which uses EP to specify the parameter τ and eliminate the parameter ξ in the standard PI and EI approaches. Consider the selection of next query point $\boldsymbol{x}_t$ as a sequential decision problem. The selection algorithm is conditioned on the last K minimum observations $y_{t-K:t-1}$ at points $\boldsymbol{x}_{t-K:t-1}$. In this context, the observations $y_{t-K:t-1}$ corresponds to the stimulations that have been experienced by the agent in the past K time. Posterior samples $\tilde{y}_{t-K:t-1}$ are drawn from the GP's posterior belief over f at points $\boldsymbol{x}_{t-K:t-1}$. Kernel density estimation (KDE) is used to estimate the probability density of the posterior samples, $\phi_{\tilde{y}_{t-K:t-1}}$. Algorithm 4 shows the steps for specifying τ by using EP. For a minimisation problem, τ is identified by applying EP to the lower quantiles of $\phi_{\tilde{y}_{t-K:t-1}}$ (see App. B.1 for further explanation). $\phi_{\tilde{y}_{t-K:t-1}}$ is analysed to identify the adaptation level, the parameters of the reward curve Φ_R and the aversion curve Φ_A in accordance with Definition 3.2, 3.3 and 3.4. However, in this application, μ_R, μ_A and the Wundt curve W are computed by negating its values as follows:

$$\mu_R = \inf\{\tilde{y} : \phi_{\tilde{y}_{t-K:t-1}}(\tilde{y}) \geq \phi_{\alpha_R}\} \tag{10}$$

$$\mu_A = \inf\ \{\tilde{y} : \phi_{\tilde{y}_{t-K:t-1}}(\tilde{y}) \geq \phi_{\alpha_A}\} \tag{11}$$

$$h_t(\tilde{y_i}) = -W(\tilde{y_i}) = -(\Phi_R(\tilde{y_i}; \mu_R, \sigma_R) - \Phi_A(\tilde{y_i}; \mu_A, \sigma_A)) \tag{12}$$

where μ_R is computed as the infimum of $100(1 - \alpha_R)$ HDR, and μ_A is computed as the infimum of $100(1 - \alpha_A)$ HDR. τ is identified to be the stimulation that invokes the maximum pleasure, i.e. $\tau = \arg\max_{\tilde{y}_{t-K:t-1}} h_t(\tilde{y_i})$.

5.3 EXPERIMENTS

We evaluated our approach with four standard test functions taken from the literature (Jamil & Yang, 2013): Branin, Six-hump camel, Goldstein-Price, Alpine-1. With Alpine-1 function, three types of dimensionality are evaluated: 3D, 5D, and 10D. All functions are typically minimisation problems, continuous, bounded and multi-modal. The key characteristics and formulae of the functions are given in App. B.2. All the experiments used Gaussian process priors for f with zero mean function

Algorithm 4 Specifying τ with EP

procedure EP($y_{t-K:t-1}$)
Input: $\alpha_R \leftarrow$ reward curve HDR, $\alpha_A \leftarrow$ aversion curve HDR
Draw posterior samples $\tilde{y}_{t-K:t-1}$ and estimate probability density $\phi_{\tilde{y}_{t-K:t-1}}$
Identify adaptation level AL at 1% HDR of $\phi_{\tilde{y}_{t-K:t-1}}$
With $\phi_{\tilde{y}_{t-K:t-1}}$, compute reward curve Φ_R parameters: μ_R, σ_R
With $\phi_{\tilde{y}_{t-K:t-1}}$, compute aversion curve Φ_A parameters: μ_A, σ_A
Compute hedonic value $h_t(\tilde{y}) = -(W(\tilde{y})) = -(\Phi_R(\tilde{y}) - \Phi_A(\tilde{y}))$
Compute $\tau = \arg\max_{\tilde{y}_{t-K:t-1}} h_t(\tilde{y})$
return τ

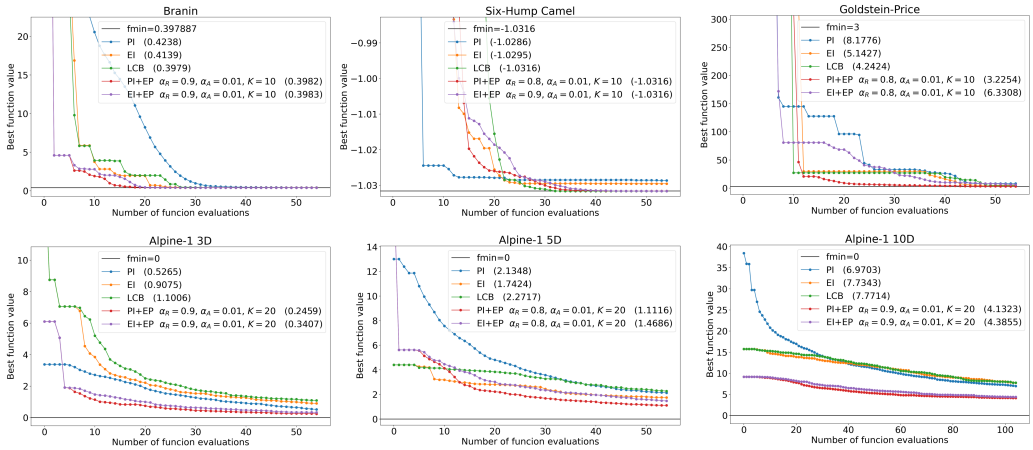

Figure 3: Comparing five acquisition methods - PI, EI, LCB, PI+EP, EI+EP - on six standard test functions.

and Matérn kernels. Each experiment was repeated for 50 trials and each trial runs 50 or 100 function evaluations. Performance was measured by the best function value observed so far.

Figure 3 compares the performance of five acquisition methods - PI, EI, LCB, PI+EP, EI+EP. PI and EI denote the standard PI and EI acquisition methods. LCB denotes the Lower Confidence Bound acquisition method (Cox & John, 1997). PI+EP and EI+EP denote our proposed methods. PI and EI used parameter $\xi = 0.01$ as suggested by Lizotte (2008). Following Cox & John (1997), LCB used parameter $\kappa = 2$. Hyperparameters used in EP are: $\alpha_R \in \{0.9, 0.8\}$, $\alpha_A = 0.01$, $K \in \{10, 20\}$. These hyperparameters are chosen following a coarse grid search over a range of values, as detailed in App. B.3. The horizontal solid black line indicates the function minimum. Figures in the bracket are the minimum function value achieved by each method. As we can observe, on Branin and Six-Hump Camel, which are easy functions, PI+EP and EI+EP performed comparably to the baseline methods. On difficult function (i.e. Goldstein-Price) and high dimensional functions (i.e. Alpine-1 3D, 5D, 10D), PI+EP outperformed other methods. Overall, our results show that by using EP to specify the parameter τ, the function landscape is being searched more effectively and this accounts for the faster convergence and better function values.

6 Conclusion

This paper proposes a method of exploration with pleasure (EP) which is formulated in accordance with "the pacer principle" and the Wundt curve from psychology. We present two demonstration implementation of EP in purely curiosity-driven RL and Bayesian optimisation. Our experiments show that EP do produce improvements in performance in both of the applications. In this paper, EP is constructed considering only one type of stimulation property. Our future work is to extend EP to handle multiple stimulation properties and study how this would improve the learning performance in curiosity-driven RL.

REFERENCES

Joshua Achiam and Shankar Sastry. Surprise-Based Intrinsic Motivation for Deep Reinforcement Learning. *arXiv e-prints*, art. arXiv:1703.01732, March 2017.

Andrew G. Barto. Intrinsic Motivation and Reinforcement Learning. In Gianluca Baldassarre and Marco Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_2.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1471–1479. Curran Associates, Inc., 2016.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Int. Res.*, 47(1):253?279, May 2013. ISSN 1076-9757.

D. E. Berlyne. Curiosity and exploration. *Science*, 153(3731):25–33, 1966. ISSN 0036-8075. doi: 10.1126/science.153.3731.25.

D E Berlyne. *Aesthetics and psychobiology.* Appleton-Century-Crofts, East Norwalk, CT, US, 1971.

Daniel E Berlyne. Curiosity and Learning. *Motivation and Emotion*, 2(2):97–175, 1978.

D.E. Berlyne. *Conflict, arousal, and curiosity*. McGraw-Hill series in psychology. McGraw-Hill, 1960.

D.E. Berlyne. Chapter 1 - the vicissitudes of aplopathematic and thelematoscopic pneumatology (or the hydrography of hedonism). In D.E. Berlyne and K.B. Madsen (eds.), *Pleasure, Reward, Preference*, pp. 1 – 33. Academic Press, 1973. ISBN 978-0-12-092550-6. doi: https://doi.org/10.1016/B978-0-12-092550-6.50006-5.

Eric Brochu, Vlad M. Cora, and Nando de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv e-prints*, art. arXiv:1012.2599, December 2010.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2019a.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019b.

Dennis D. Cox and Susan John. Sdo: A statistical method for global optimization. In *in Multidisciplinary Design Optimization: State-of-the-Art*, pp. 315–329, 1997.

William N Dember and Robert W Earl. Analysis of exploratory, manipulatory, and curiosity behaviors., 1957.

R.E. Franken. *Human Motivation*. Thomson/Wadsworth, 2006. ISBN 9780495090816.

Justin Fu, John Co-Reyes, and Sergey Levine. Ex2: Exploration with exemplar models for deep reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2577–2587. Curran Associates, Inc., 2017.

Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1109–1117. Curran Associates, Inc., 2016.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50 (2):120–126, 1996. ISSN 00031305.

Momin Jamil and Xin-She Yang. A Literature Survey of Benchmark Functions For Global Optimization Problems. *arXiv e-prints*, art. arXiv:1308.4008, August 2013.

Celeste Kidd and BenjaminY. Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88 (3):449 – 460, 2015. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2015.09.010.

Celeste Kidd, Steven T. Piantadosi, and Richard N. Aslin. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLOS ONE*, 7 (5):1–8, 05 2012. doi: 10.1371/journal.pone.0036399.

Youngjin Kim, Wontae Nam, Hyunwoo Kim, Ji-Hoon Kim, and Gunhee Kim. Curiosity-bottleneck: Exploration by distilling task-specific novelty. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3379–3388, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Dennis K. Kinney and Jerome Kagan. Infant attention to auditory discrepancy. *Child Development*, 47(1):155–164, 1976. ISSN 00093920, 14678624.

H. J. Kushner. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, 03 1964. ISSN 0021-9223. doi: 10.1115/1.3653121.

Kristjan Laane. *The Ins and Outs of Pleasure: Roles and Importance of Hedonic Value*. PhD thesis, University of Cambridge, 2011.

Daniel James Lizotte. *Practical Bayesian Optimization*. PhD thesis, University of Alberta, CAN, 2008. AAINR46365.

Kathryn E. Merrick. Novelty and beyond: Towards combined motivation models and integrated learning architectures. In Gianluca Baldassarre and Marco Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 209–233. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_9.

Peter M. Milner. Brain-stimulation reward: A review. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(1):1–36, 1991. ISSN 0008-4255. doi: 10.1037/h0084275.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518, 2015. doi: 10.1038/nature14236.

J. Močkus. On bayesian methods for seeking the extremum. In G. I. Marchuk (ed.), *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pp. 400–404, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg. ISBN 978-3-540-37497-8.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2125–2133. Curran Associates, Inc., 2015.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2721?2730. JMLR.org, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 2778–2787, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006. ISBN 026218253X.

Rob Saunders. *Curious Design Agents and Artificial Creativity A Synthetic Approach to the Study of Creative Behaviour*. PhD thesis, Department of Architectural and Design Science, Faculty of Architecture, University of Sydney, Australia, February 2002.

Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphael Marinier, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference on Learning Representations*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv e-prints*, art. arXiv:1707.06347, Jul 2017.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

Paul J Silvia. Curiosity and motivation. In *The Oxford handbook of human motivation.*, Oxford library of psychology., pp. 157–166. Oxford University Press, New York, NY, US, 2012. ISBN 978-0-19-539982-0 (Hardcover).

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. *arXiv e-prints*, art. arXiv:1507.00814, July 2015.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, Sep 2012. ISSN 1611-7530. doi: 10.1007/s12064-011-0142-z.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2753–2762. Curran Associates, Inc., 2017.

M. Zuckerman. *Sensation Seeking: Beyond the Optimal Level of Arousal*. Psychology Revivals. Taylor & Francis Group, 2016. ISBN 9781848727793.

A Application in Purely Curiosity-Driven RL

A.1 Purely Curiosity-Driven RL Algorithm

Alg. 5 shows the steps of purely curiosity-driven RL with the intrinsic rewards generated by using EP. We use the same algorithm as reported in Burda et al. (2019a), with the only difference being that we compute the intrinsic reward by using EP.

A.2 Surprisingness density of Atari Games

The four Atari games used in our experiments are Pong, Riverraid, SpaceInvaders and Asterix. These four games were chosen because the surprisingness experienced by the agent in these games exhibit distinguished patterns of density. Figure 4 shows the density plot of the surprisingness contained in the surprisingness buffer U at time steps 1e6, 2e6, 3e6 and 4e6.

A.3 RL Hyperparameters

For the application of EP in purely curiosity-driven RL, all of the experiments used the Proximal Policy Optimisation (PPO) RL algorithm (Schulman et al., 2017). The embedding network and policy networks had identical architectures and were based on the standard convolutional networks used in Atari experiments as suggested in Mnih et al. (2015). The feature layer in the embedding network had a dimension of 512. A learning rate of 0.0001 was used for all networks. All experiments used 32 parallel environments, rollouts of length 128, three optimisation epochs per rollout. For pre-processing, all game images were converted to grayscale and resized to to size 84×84. All agents' policy, value network and forward dynamics functions used a stack of 4-frame historical observation $[x_{t-3}, x_{t-2}, x_{t-1}, x_t]$.

Algorithm 5 Purely curiosity-driven RL with EP

Input: $N \leftarrow$ number of rollouts, $J \leftarrow$ length of rollout
Input: $N_{opt} \leftarrow$ number of optimisation steps, $N_{ts} \leftarrow$ number of time steps
Initialize: $t = 0$, double-ended buffer $U = \{\}$ of size K
Sample state $s_0 \sim p_0(s_0)$
for $n = 1$ **to** N **do**
  for $j = 1$ **to** J **do**
    Sample $a_t \sim \pi(a_t|s_t)$
    Sample $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    Quantify surprisingness u_t
    Compute intrinsic reward $r_t^i = \text{EP}(U, u_t)$
    Add s_t, s_{t+1}, a_t, u_t to optimisation batch B_n
    $t = t + 1$
  Add $\{u\}_{j=1}^{J}$ to buffer U
  Normalise the intrinsic rewards contained in B_n
  Calculate returns R_n and advantages A_n for intrinsic rewards
  for $j = 1$ **to** N_{opt} **do**
    Optimise θ_π wrt PPO loss on batch B_n, R_n, A_n using Adam
  if $t > N_{ts}$ **then**
    break

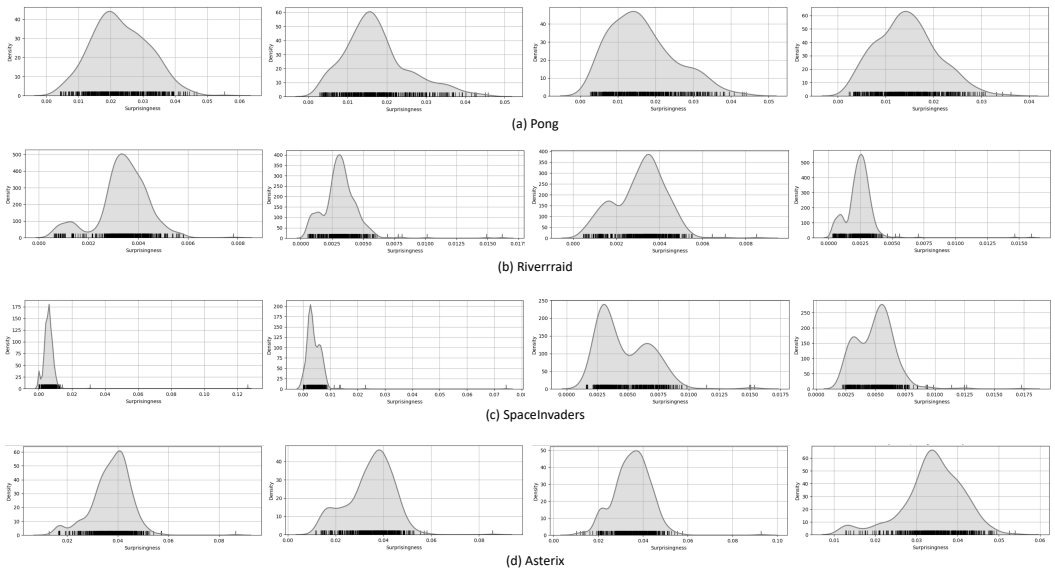

Figure 4: Density plots of surprisingness contained in the surprisingness buffer U at time steps 1e6, 2e6, 3e6, 4e6 (from left to right) in Pong (first row), Riverrraid (second row), SpaceInvaders (third row) and Asterix (fourth row).

Prediction-error-based surprise quantification method is used in the experiments. State s_t is encoded into a feature vector $v(s_t)$ by using a fix, randomly initialised convolutional network, which has been reported to perform well in purely curiosity-driven RL by Burda et al. (2019a;b). Given the current encoded state $v(s_t)$ and action a_t, a forward dynamics network is used to predict the encoded next state $\hat{v}(s_{t+1})$. Surprisingness u_t is computed as the prediction error of the forward dynamics network, i.e. the error between the predicted next state $\hat{v}(s_{t+1})$ and the ground truth next state $v(s_{t+1})$.

The implementation of PPO, feature network and forward dynamics network are adopted from the codes released by Burda et al. (2019a;b)[1] [2].

A.4 EP HYPERPARAMETERS

Ablations for reward HDR α_R and aversion HDR α_A. The best combination of α_R and α_A is selected by a grid search. The effects of α_R and α_A were examined with a range of values as follows: $\alpha_R \in \{0.9, 0.8\}$, $\alpha_A \in \{0.05, 0.03, 0.01\}$. We performed the ablations and analysed the performance of the EP agent on four Atari games: Pong, Riverraid, SpaceInvaders and Asterix. Figure 5 shows the learning curves of different variants of EP. All learning curves are the average of three runs; confidence intervals are omitted in the figure for clearer presentation. Tab. 3 shows the results for all the ablations we performed on 4 games. On Pong, the densities of the surprisingness history are consistent unimodal distributions (see Figure 4 (a)). In this game, an EP agent with a narrow Wundt curves as specified by $\alpha_R = 0.8$ and $\alpha_A = 0.05$ performed particularly well. In contrast, the surprisingness densities in Riverraid and SpacceInvaders are skewed, sharply peaked bimodal distributions (see Figure 4 (b), (c)), which are not able to be handled well by EP. As we can observe, the improvement produced by EP on these two games are less pronounced. On Riverraid, EP with $\alpha_R = 0.8$, $\alpha_A = 0.01$ performed the best. On SpaceInvaders, all variants of EP showed unstable performance; the best learning curve was yielded by EP with $\alpha_R = 0.9$, $\alpha_A = 0.03$. On Asterix, all variants of EP performed well with the best learning curve being achieved by EP with $\alpha_R = 0.9$ and $\alpha_A = 0.03$.

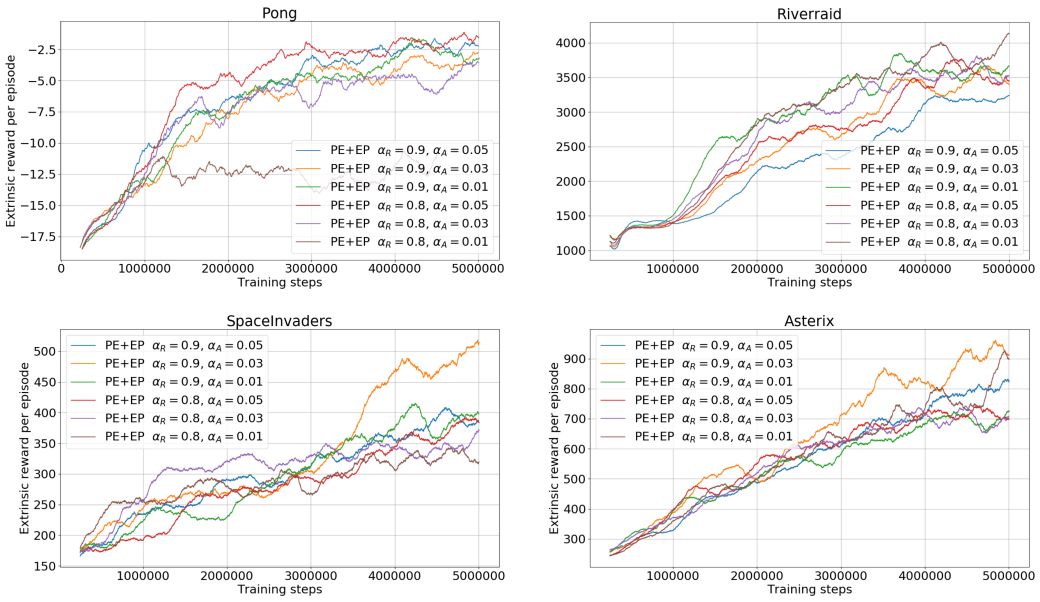

Figure 5: Ablation results for reward HDR α_R and aversion HDR α_A, with $\alpha_R \in \{0.9, 0.8\}$, $\alpha_A \in \{0.05, 0.03, 0.01\}$, as evaluated on Pong, Riverraid, SpaceInvaders and Asterix.

Ablations for surprisingness buffer size K. The effect of the size of surprisingness buffer K was examined with values $\{128, 256, 512, 1024\}$. With a rollout of length 128, a buffer size of 128, 256, 512, 1024 means storing the last one, two, four and eight rollouts of surprisingness, respectively. We performed ablations and analysed the performance of the EP agent on two Atari games: Pong and Riverraid. Results are shown in Figure 6. The results indicate that increasing the buffer size from $K = 128$ to $K = 512$ produced increasingly better learning curves. But further increase of buffer

[1] https://github.com/openai/large-scale-curiosity.

[2] https://github.com/openai/random-network-distillation

Table 2: Ablation results on Pong, Riverraid, SpaceInvaders and Asterix.

Method	Pong	Riverraid	SpaceInvaders	Asterix
PE	-8.8±1.8	2791.4±562.8	294.8±48.7	551.3±124.3
PE+EP - $\alpha_R = 0.9, \alpha_A = 0.05$	-3.7±1.8	3242.6±216.8	415.1±110.6	795.4±144.5
PE+EP - $\alpha_R = 0.9, \alpha_A = 0.03$	-2.9±1.9	3452.2±475.8	**523.1±82.8**	**912.9±210.6**
PE+EP - $\alpha_R = 0.9, \alpha_A = 0.01$	-2.4±1.7	3764.0±377.6	415.6±74.5	752.6±164.7
PE+EP - $\alpha_R = 0.8, \alpha_A = 0.05$	**-1.1±1.6**	3289.1±580.8	351.7±66.5	705.4±175.4
PE+EP - $\alpha_R = 0.8, \alpha_A = 0.03$	-2.7±1.5	3403.9±613.9	368.0±69.4	761.6±130.1
PE+EP - $\alpha_R = 0.8, \alpha_A = 0.01$	-8.8±3.7	**4215.3±383.5**	318.8±66.7	804.2±169.7

size to $K = 1024$ did not yield any better performance. The optimal choice for K is 512. The results show that EP does not require a large buffer of surprisingness history to work well.

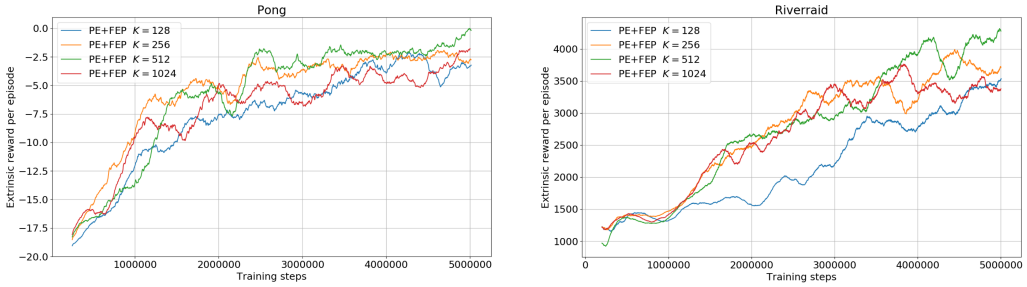

Figure 6: Ablation results for surprisingness buffer size K, with $K \in \{128, 256, 512, 1024\}$, as evaluated on Pong and Riverraid,.

B Application in Bayesian Optimisation

B.1 EP in Minimisation Problem

Figure 7 shows an example of using EP to specify the parameter τ. For a minimisation problem, τ is identified by applying EP to the lower quantiles of $\phi_{\tilde{y}_{t-K:t-1}}$.

B.2 Standard Test Functions

The application of EP in BO was evaluated on four standard test functions: Branin, Six-hump camel-back, Goldstein-Price and Alpine-1. The formulae of the functions are as follows.

Branin. This function is two-dimensional given by:

$$f_{BR}(x_1, x_2) = (x_2 - (\frac{5.1}{4\pi^2})x_1^2 + (\frac{5}{\pi})x_1 - 6)^2 + 10(1 - \frac{1}{8\pi})\, cos(x_1) + 10 \tag{13}$$

$$x_1 \in [-5, 10] \tag{14}$$

$$x_2 \in [0, 15] \tag{15}$$

The function has three identical global minima of 0.397887 at $\mathbf{x} = (-3.1416, 12.275)$, $\mathbf{x} = (3.1416, 2.275)$ and $\mathbf{x} = (9.42478, 2.475)$. The Branin function is an easy optimisation as the function has three identical global minima, each of which lies in a wide and shallow basin.

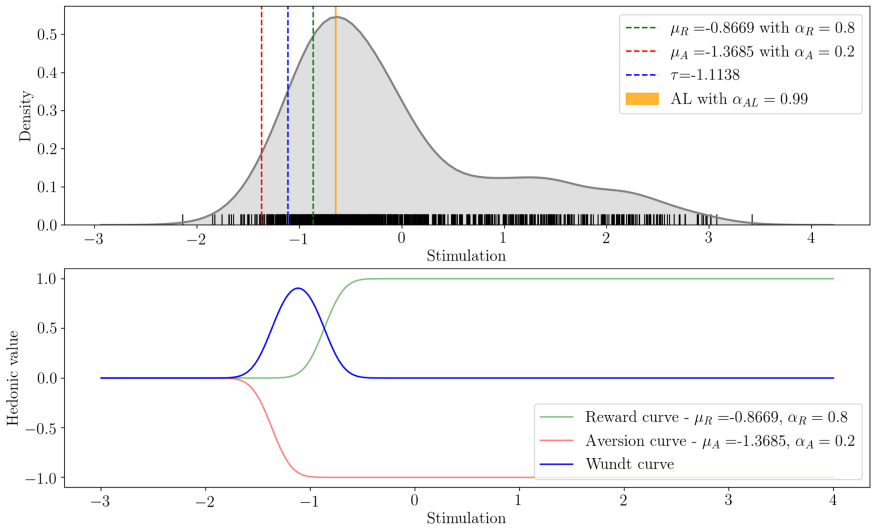

Figure 7: For a minimisation problem in BO, τ is identified by applying EP to the lower quantiles of $\phi_{\tilde{y}_{t-K:t-1}}$.

Six-hump camel-back. This function takes a two-dimensional input given by:

$$f_{6H}(x_1, x_2) = (4 - 2.1x_1^2 + x_2^3) \cdot x_1^2 + x_1x_2 + (-4 + 4x_2^2) \cdot x_2^2 \tag{16}$$
$$x_1 \in [-2, 2] \tag{17}$$
$$x_2 \in [-1, 1] \tag{18}$$

The function has two identical global minima of -1.0316 at $\mathbf{x} = (-0.0898, 0.7126)$ and $\mathbf{x} = (0.0898, -0.7126)$.

Goldstein-Price. This is a two-dimensional function of the form:

$$f_{GP}(x_1, x_2) = (1 + (x_1 + x_2 + 1)^2 \, g_1(x_1, x_2)) \cdot (30 + (2x_1 - 3x_2)^2 \, g_2(x_1, x_2)) \text{ where} \tag{19}$$
$$g_1(x_1, x_2) = 19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2 \tag{20}$$
$$g_2(x_1, x_2) = 18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2 \tag{21}$$
$$x_1, x_2 \in [-2, 2] \tag{22}$$

The global minimum of this function is 3.0 at $\mathbf{x} = (0, -1)$. The Goldstein-Price function presents a more difficult optimisation because the function's single minimum lies in a small convex basin surrounded by steep sides (Lizotte, 2008).

Alpine N. 1 function. This function is with n-dimensional input given by:

$$f_{A1}(x_1, \ldots, x_n) = \sum_{i=1}^{n} |x_i \, sin(x_i) + 0.1x_i| \tag{23}$$
$$x_n \in [0, 10] \tag{24}$$

This function can be defined on any positive input domain but it is usually evaluated on $x_i \in [0, 10]$ for $i = 1, \ldots, n$. The function has a global minimum of 0 located at $\mathbf{x} = (0, \ldots, 0)$.

Table 3 summarises the key characteristics of the standard test functions.

Table 3: Key characteristics of the standard test functions: Branin, Six-hump camel-back, Goldstein-Price and Alpine-1.

Test function	No of dimensions	No of global minima	Global minima
Branin	2	3	0.397887
Six-hump camel	2	2	-1.0316
Goldstein and Price	2	1	3.0
Alpine-1 3D	3	1	0
Alpine-1 5D	5	1	0
Alpine-1 10D	10	1	0

Table 4: Ablation results on Branin, Six-hump Camel-back and Goldstein-Price.

Method	**Branin (0.397887)**	**Six-hump Camel (-1.0316)**	**Goldstein Price (3.0)**
PI	0.4238±0.0231	-1.0286±0.0006	8.178±2.952
EI	0.4139±0.0042	-1.0295±0.0010	5.143±1.125
LCB	**0.3979±0.0000**	**-1.0316±0.0000**	4.242±1.157
PI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 10$	0.3982±0.0005	**-1.0316±0.0000**	3.584±1.048
PI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 20$	0.3988±0.0011	-1.0310±0.0010	4.614±2.300
PI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 10$	0.3982±0.0006	**-1.0316±0.0000**	**3.225±0.262**
PI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 20$	0.3985±0.0010	-1.0298±0.0020	6.746±6.490
PI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 10$	0.3982±0.0004	-1.0315±0.0003	4.030±1.977
PI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 20$	0.3990±0.0017	-1.0300±0.0023	7.907±7.938
EI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 10$	0.3983±0.0006	**-1.0316±0.0000**	11.733±6.389
EI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 20$	0.3985±0.0013	-1.0313±0.0008	7.054±3.407
EI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 10$	0.3983±0.0007	-1.0316±0.0001	6.331±3.542
EI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 20$	0.3990±0.0019	-1.0283±0.0030	13.355±8.791
EI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 10$	0.3987±0.0009	-1.0315±0.0003	10.509±2.681
EI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 20$	0.3996±0.0029	-1.0306±0.0022	4.438±1.159

B.3 EP HYPERPARAMETERS

EP has three hyperparameters: the reward HDR α_R, the aversion HDR α_A and the history buffer size K. The effects of these hyperparameters were examined with the six standard test functions. A coarse grid search was done with a range of values as follows: $\alpha_R \in \{0.9, 0.8, 0.7\}$, $\alpha_A \in \{0.01\}$, $K \in \{10, 20\}$. Figures 8 shows the results of the ablations. PI+EP denotes BO using the PI acquisition function with EP. EI+EP denotes BO using the EI acquisition function with EP. The horizontal solid black line indicates the function minimum. Each curve is the averaged results over 50 trials. Tables 4 and 5 list the best function values achieved by the PI, EI, LCB and different variants of PI+EP, EI+EP as evaluated on six standard test functions. We observed that EPs with $\alpha_R \in \{0.9, 0.8\}$, $\alpha_A \in \{0.01\}$, $K = 10$ performed consistently well on low dimensional functions as can be seen in Branin, Six-Hump Camel and Goldstein-Price. For higher dimensional functions, i.e. Alpine-1 3D, 5D and 10D, EP requires a larger history buffer with $K = 20$ to perform well.

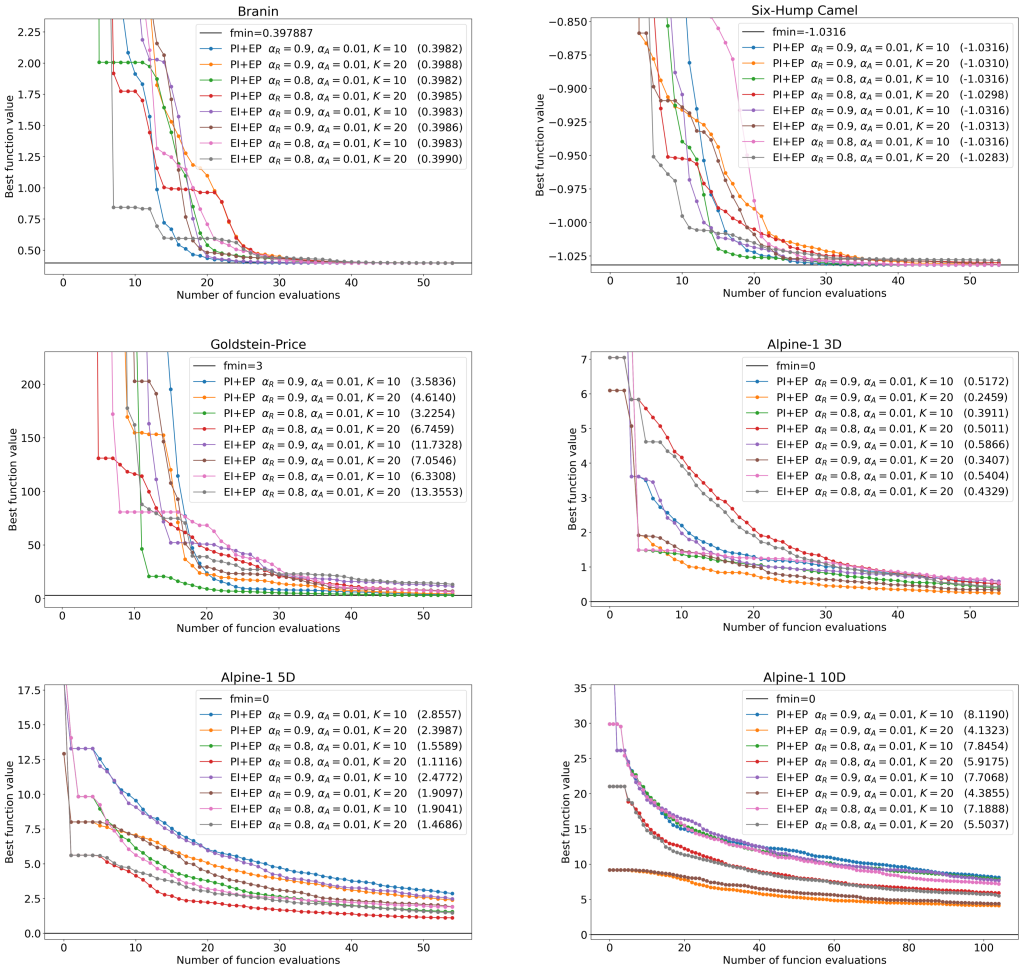

Figure 8: Effects of reward HDR $\alpha_R \in \{0.9, 0.8\}$, aversion HDR $\alpha_A \in \{0.01\}$ and history buffer size $K \in \{10, 20\}$ on the performance of PI+EP and EI+EP as evaluated on six standard test functions.

Table 5: Ablation results on Alpine-1 3D, Alpine-1 5D and Alpine-1 10D.

Method	**Alpine-1 3D (0.0)**	**Alpine-1 5D (0.0)**	**Alpine-1 10D (0.0)**
PI	0.527±0.435	2.135±1.037	6.970±2.979
EI	0.907±0.529	1.742±0.880	7.734±2.280
LCB	1.101±0.628	2.272±1.019	7.771±2.977
PI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 10$	0.517±0.341	2.856±1.813	8.119±2.921
PI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 20$	**0.246±0.161**	2.399±1.116	**4.132±1.781**
PI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 10$	0.391±0.249	1.559±1.004	7.845±2.749
PI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 20$	0.501±0.339	**1.111±0.606**	5.917±2.252
PI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 10$	0.473±0.352	1.999±1.206	7.122±2.428
PI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 20$	0.364±0.328	2.442±1.259	6.925±3.371
EI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 10$	0.587±0.319	2.477±1.388	7.707±3.758
EI+EP - $\alpha_R = 0.9$, $\alpha_A = 0.01$, $K = 20$	0.341±0.228	1.909±1.108	4.385±1.661
EI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 10$	0.540±0.382	1.904±0.853	7.189±2.696
EI+EP - $\alpha_R = 0.8$, $\alpha_A = 0.01$, $K = 20$	0.433±0.338	1.468±0.826	5.503±2.054
EI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 10$	0.596±0.357	2.084±1.019	6.243±2.896
EI+EP - $\alpha_R = 0.7$, $\alpha_A = 0.01$, $K = 20$	0.273±0.292	2.389±1.386	7.339±3.052