An information-theoretic perspective on intrinsic motivation in reinforcement learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The standard reinforcement learning (RL) framework faces the problem of transfer learning and sparse rewards explorations. To address these problems, a large number of heterogeneous intrinsic motivation have been proposed, like reaching unpredictable states or unvisited states. Yet, it lacks a coherent view on these intrinsic motivations, making hard to understand their relations as well as their underlying assumptions. Here, we propose a new taxonomy of intrinsic motivations based on information theory: we computationally revisit the notions of surprise, novelty and skill learning and identify their main implementations through a short review of intrinsic motivations in RL. Our information theoretic analysis paves the way towards an unifying view over complex behaviors, thereby supporting the development of new objective functions.

1 Introduction

Intrinsic motivation (IM) aims to describe to natural drive of children to explore their environment Ryan and Deci (2000) and acquire new skills Piaget and Cook (1952). Simply stated, intrinsic motivation Barto (2013) is about doing something for its inherent satisfaction rather than to get a positive feedback from the environment Ryan and Deci (2000). Taking inspiration from the psychological concept, numerous intrinsic motivations have been introduced in RL: an agent learns by trials and errors to maximize its expected discounted cumulative intrinsic rewards. They show tremendous improvements on sparse rewards tasks and transfer learning with respect to standard RL methods Eysenbach et al. (2018); Bellemare et al. (2016). For example, one can incite an agent to reach unvisited states Bellemare et al. (2016).

In practice, there is no consensus on a qualitative definition of intrinsic motivations, such that one can name “intrinsic" any task-agnostic rewards. Two issues result from this absence of a common theory: 1- there are a plethora of heterogeneous IMs; 2- the core mechanisms that underpin their success is unclear. In practice, it makes hard to identify relevant avenues of works.

Here, we take a step back and bound important results to information theoretic objectives. We identify three important information theoretic objectives and relate them to psychological concepts as well as the current deep RL literature. First, we define the surprise as the expected information gain over true forward models and highlight that, under harmful approximations, it encompasses intrinsic rewards based on the expected information gain, prediction error or learning progress. Second, we revisit novelty as actively learning representations and exhibit that it is currently maximized through two line of works, one based on variational inference and one with k-nearest neighbors. Third, we introduce the skill learning objective and find that either discriminator based or states-goals based methods maximize it. We expect our computational taxonomy to play a crucial role in understanding the principles that underlies complex behaviors, and thus to help building intrinsically motivated agents.

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Surprise: $I(S'; \Phi_T \| h, S, A)$			
Formalism	IG over forward model	Prediction error	Learning progress
Rewards	$D_{KL}(p(\Phi\|h, s, a, s')\|\|p(\Phi\|h))$	$\|\|s' - \hat{s}'\|\|_2^2$	$\Delta\|\|s' - \hat{s}'\|\|_2^2$
Novelty: $I(S; Z)$			
Formalism	K-nearest neighbors	Variational inference	
Rewards	$-\log \|\|s' - nn_k(S_b, s')\|\|_2$	$-\log q_d(s'\|z) + D_{KL}(q_e(z\|s')\|\|p(z))$	
Skill learning: $I(G; u(\mathcal{T}))$			
Formalism	Fixed goal distribution	Goal-state achievement	Diverse goals
Rewards	$\log p(g\|s')$	$-\|\|s_g - s'\|\|_2^2$	$(1 + \alpha_{skew}) \log p(s_g)$

Table 1: Summary of our taxonomy of intrinsic motivations in DRL. Please, refer to the corresponding sections for more details about methods and notations.

2 Background

Reinforcement learning. In a Markov Decision Process (MDP), an agent interacts with an unknown environment in the following way: the agent gets a state $s_0 \in S$ that follows an initial state distribution $p_0(s)$. Its policy π selects actions $a \in A$ to execute depending on its current state $s \in S$, $a \sim \pi(\cdot|s)$. Then, a new state is returned according to the transition dynamics ϕ_T, $s' \sim p(\cdot|s, a, \phi_T)$. The agent repeats the procedure until it reaches a particular state or exceeds a fixed number of steps T. In RL, the agent learns π to maximize the expected cumulative discounted reward given at each time step by the reward function $R(s_t, a_t, s_{t+1}$. In goal-conditioned RL, the reward function and the policy π^g both depends on a given goal.

Information theory. Here, we provide the basics about information theory. The Shannon entropy quantifies the mean necessary information to determine the value of a random variable. Let X be a random variable with a law of density $p(X)$ satisfying the normalization and positivity requirements, we define its entropy by $H(X) = -\int_X p(x) \log p(x) dx$.

The mutual information allows to quantify the information contained in a random variable X about an other random variable Y and is defined by $I(X; Y) = H(X) - H(X|Y)$.

3 Three information theoretic objectives

In this section, we explain our three information theoretic objectives and shortly review the mains approaches they encompass. Table 1 sums up our findings.

3.1 Surprise

In this section, we assume the agent learns either a forward model of the environment parameterized by $\phi \in \Phi$. The forward model computes the next-state distribution conditioned on a tuple state-action $p(S'|S, A, \phi)$. Typically, ϕ can be the parameters of a neural network. Trying to approximate the true model, the agent maintains an approximate distribution $p(\Phi|h)$ of models, where $h_t = h$ refers to the ordered history of interactions $((s_0, a_0, s_1), \ldots, (s_{t-1}, a_{t-1}, s_t))$.

In this paper, we adopt the definition of Itti and Baldi (2009), which defines surprise as the discrepancy between a prior distribution of beliefs and the posterior probability distribution following an observation. In our formalism, we assume that there is a distribution of true models $p(\Phi_T)$ that underpins the transition function of the environment T. In contrast with Φ, this is a property of the environment. One can see this distribution as a Dirac distribution if only one model exists or as a categorical distribution of several forward models. We define the expected information gain over the true models:

$$IG(h, A, S', S, \Phi_T) = I(S'; \Phi_T|h, A, S) = H(\Phi_T|h, A, S) - H(\Phi_T|h, A, S, S') \tag{1a}$$

$$= \mathbb{E}_{(s,a)\sim p(\cdot|h),\, \phi_T \sim p(\cdot), s' \sim p(\cdot|s,a,\phi_T)} \log p(s'|s, a, h, \phi_T) - \log p(s'|s, a, h). \tag{1b}$$

We found three approximations of our formalism of surprise.

Surprise as prediction error. First, we can assume that the true forward model is a deterministic forward model, *i.e* $\log p(s'|s,a,h,\phi_T) = Const$. Then, assuming a Gaussian probability distribution, we end up with (cf. Appendix A.1):

$$IG(h, A, S', S, \Phi_T) \propto \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h),\, s'\sim p(\cdot|s,a,\phi_T) \\ \phi\sim p(\cdot|h),\, \phi_T\sim p(\cdot)}} ||s' - \hat{s}'||_2^2 + Const \tag{2}$$

where $\hat{s}'$ is the prediction of state following (s, a). Several works implement such objective using the ground state space or a learnt representation Ermolov and Sebe (2020); Pathak et al. (2017); Kim et al. (2019). Despite simplicity, their underlying assumption makes them sensitive to stochastic transitions Burda et al. (2019).

Surprise as learning progress. One can also relax the determinism constraint and set

$$\log p(s'|s,a,h,\phi_T) - \log p(s'|s,a,h) \approx \log p(s'|s,a,h') - \log p(s'|s,a,h) = \Delta||s' - \hat{s}'||_2^2 \tag{3}$$

where h' concatenates h with an arbitrary number of additional interactions (cf. Appendix A.2). Such loss is called the *learning progress* and has also be used several time in the literature, in particular because it is more robust to stochastic transitions than prediction errors Schmidhuber (1991); Oudeyer et al. (2007); Kim et al. (2020). However, we are not aware of an application to complex RL environments.

Surprise as prediction disagreement. By assuming $\Phi \approx \Phi_T$, it's also possible to approximate the surprise by a form of prediction disagreement:

$$IG(h, A, S', S, \Phi_T) \approx \mathbb{E}_{\substack{\phi\sim p(\cdot|h,s,a,s'),\, \phi_T\sim p(\cdot) \\ (s,a)\sim p(\cdot|h), s'\sim p(\cdot|s,a,h,\phi_T)}} \log p(s'|s,a,h,\phi) - \log p(s'|s,a,h) \tag{4a}$$

$$= \mathbb{E}_{\substack{\phi\sim p(\cdot|h,s,a,s'),\, \phi_T\sim p(\cdot) \\ (s,a)\sim p(\cdot|h), s'\sim p(\cdot|s,a,h,\phi_T)}} \log p(s'|s,a,h,\phi) - \log \sum_{\phi\in\Phi} p(s'|\phi,h,s,a)p(\phi|h). \tag{4b}$$

where we simply marginalize out the ϕ in the right-hand term. This amounts to compute the difference of prediction among the forward models. Similar equations have been investigated in Shyam et al. (2019); Pathak et al. (2019); Yao et al. (2021); Sekar et al. (2020).

Surprise as expected information gain. By extending $\Phi \approx \Phi_T$ to the expectation part, we can further increase the approximation to match the expected information gain Sun et al. (2011), used as an intrinsic motivation Houthooft et al. (2016); Achiam and Sastry (2017):

$$IG(h, A, S', S, \Phi_T) = \mathbb{E}_{(s,a)\sim p(\cdot|h), s'\sim p(\cdot|s,a,h,\phi_T)} D_{KL}(p(\Phi|h,s,a,s')||p(\Phi|h)) \tag{5}$$

The last two approaches are computationally expensive as they require either 1- several forward models or 2- having a tractable probability distribution over forward models. We considering a density of a model rather than a forward model, Equation 5 amounts to maximize the pseudo-count Martin et al. (2017); Ostrovski et al. (2017) (cf. Bellemare et al. (2016)).

3.2 Novelty

Barto et al. (2013) defend that *an observation is novel when a representation of it is not "close enough" to any representation found in memory*. Following this definition, we propose to formalize novelty seeking behaviors as those that actively maximize $I(S;Z) = H(S) - H(S|Z)$ where Z is a low-dimensional space ($|Z| \leq |S|$). This objective is commonly known as the *infomax* principle Linsker (1988); in our case, it amounts to actively learning a representation of the environment. Most of works focus on actively maximizing the entropy of state distribution while a representation learning function minimizes $H(S|Z)$. Furthermore, if one assumes that $Z = S$, the infomax principle collapses to an entropy maximization $H(S)$. Thus, we will focus on the maximization of $H(S)$.

Novelty with variational inference The most evident way to maximize the entropy of states consists in maximizing $H(\rho(s))$ where $\rho(s) = p(s|\rho)$ approximates the state-visitation probability distribution. But computing $\rho(s)$ is challenging in high-dimensional state spaces. Several methods propose to estimate $\rho(s)$ using variational inference Zhang et al. (2021); Islam et al. (2019); Lee et al. (2019); Pong et al. (2020) based on variational autoencoder architectures. Assuming z is a compressed latent variable, $p(z)$ a prior distribution and q_d a neural network that ends with a Gaussian with a unit-diagonal co-variance matrix,:

$$\log \rho(s') \gtrapprox -\log q_d(s'|z) + D_{KL}(q_e(z|s')||p(z)) = R(s,a,s') \tag{6}$$

Novelty with k-nearest neighbors It is also possible to approximate the entropy of a distribution using samples and their k-nearest neighbors (k-nn) Singh et al. (2003); Kraskov et al. (2004). Assuming $nn_k(S_b, s_i)$ is a function that outputs the k-th closest state to s_i in the set S_b, the unbiased approximation can be written as:

$$-\log \rho(s') \propto -\log ||s' - nn_k(S_b, s')||_2 = R(s, a, s'). \tag{7}$$

Intuitively, it means that the entropy is proportional to the average distance between states and their neighbors. Few methods provably relates to such estimations, but several approaches take advantage of the distance between state and neighbors to generate intrinsic rewards Tao et al. (2020); Yarats et al. (2021); Seo et al. (2021). In practice, k-nn methods are simpler than variational methods as they do not require to fit a model. They achieve SOTA results on the hard exploration task Montezuma's revenge Bougie and Ichise (2020); Badia et al. (2019)

3.3 Skill learning

In our everyday life, nobody has to think about having to move his arms' muscles to grasp an object. A command to take the object is just issued. This can be done because an acquired skill can be effortlessly reused. Skill abstraction denotes the ability of an agent to learn a representation of diverse skills. We formalize skill learning as maximizing the mutual information between the goal $g \in G$ and some of the rest of the contextual states $u(\tau) \in u(\mathcal{T})$, denoted as $I(G; u(\mathcal{T}))$ where $\tau \in \mathcal{T}$ is a trajectory and u a function that extracts a subpart of the trajectory (last state for example). Most of works maximize $I(G; S) = H(G) - H(G|S)$ so that we will refer to this objective.

Skill learning with a discriminator The first approach Florensa et al. (2017); Eysenbach et al. (2018); Zhang et al. (2020); Baumli et al. (2021); Sharma et al. (2020); Choi et al. (2021); Aubret et al. (2020); Hansen et al. (2020) assumes the goal space is arbitrarily provided except for the semantic meaning of a goal. In this setting, the agent samples goals uniformly from G (maximal $H(G)$), and it progressively assigns all possible goals to a part of the state space. To do this assignment, the agent learns a discriminator that approximates $q(g|s) \approx p(g|s)$ and maximizes with q and the its skills:

$$-H(G|S) = \mathbb{E}_{g \sim p(g), s \sim \pi^g} \log q(g|s),\ R((s_g, s, a, s') = \log q(g|s) \tag{8}$$

such that each skill strives to reach the area assigned to it by the discriminator.

Skill learning as reaching states-goals Other set of works consider that the goal space is the state space ($G = S_g$) and separately maximize $-H(S_g|S)$ and $H(S_g)$. Let's start with the first term. If $\log p(s_g|s')$ is modelled as an unparameterized Gaussian with a unit-diagonal co-variance matrix, we have $\log p(s_g|s') \propto -||s_g - s'||_2^2 + Const$ so that one can reward an agent according to Levy et al. (2019); Zhao et al. (2019); Nachum et al. (2018); Kim et al. (2021); Nair et al. (2018); Li et al. (2021b) (where s can also be replaced by a learnt representation):

$$R(s_g, s, a, s') = -||s_g - s'||_2^2. \tag{9}$$

Let us now consider the maximization of $H(S_g)$. It can be maximized by over-sampling low-density state-goals Warde-Farley et al. (2019); Pitis et al. (2020); Zhao and Tresp (2019); Aubret et al. (2021); Li et al. (2021a). This can be formalized Pong et al. (2020) as having a high-level agent that samples goals to maximize:

$$R(s_g) = (1 + \alpha_{skew}) \log p(s_g) \tag{10}$$

where $\alpha_{skew} < 0$ determines the weight of low-density state-goals.

4 Conclusion

Is there an information theoretic principle that supports the emergence of complex behaviors ? Our analysis supports that crucial behaviors like exploration and skill learning are explained by three information theoretic objectives maximization. We argued for the relation between our unifying objective to notions from psychology like novelty, surprise or skill learning. As such, we expect that our results will favor the design of new information theoretic open-ended learners. Future works may refine our *Skill learning* objective, tackle the problem of unifying our three objectives or may try to encompass more works like bottleneck research (McGovern and Barto, 2001).

References

Joshua Achiam and Shankar Sastry. 2017. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732* (2017).

Arthur Aubret, Salima Hassas, et al. 2021. DisTop: Discovering a Topological representation to learn diverse and rewarding skills. *arXiv preprint arXiv:2106.03853* (2021).

Arthur Aubret, Laëtitia Matignon, and Salima Hassas. 2020. ELSIM: End-to-end learning of reusable skills through intrinsic motivation. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, et al. 2019. Never Give Up: Learning Directed Exploration Strategies. In *International Conference on Learning Representations*.

Andrew Barto, Marco Mirolli, and Gianluca Baldassarre. 2013. Novelty or surprise? *Frontiers in psychology* 4 (2013), 907.

Andrew G Barto. 2013. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*. Springer, 17–47.

Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. 2021. Relative Variational Intrinsic Control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6732–6740.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*. 1471–1479.

Nicolas Bougie and Ryutaro Ichise. 2020. Skill-based curiosity for intrinsically motivated reinforcement learning. *Machine Learning* 109, 3 (2020), 493–512.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. 2019. Large-Scale Study of Curiosity-Driven Learning. In *International Conference on Learning Representations*.

Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. 2021. Variational Empowerment as Representation Learning for Goal-Conditioned Reinforcement Learning. In *International Conference on Machine Learning*. PMLR, 1953–1963.

Aleksandr Ermolov and Nicu Sebe. 2020. Latent World Models For Intrinsically Motivated Exploration. *Advances in Neural Information Processing Systems* 33 (2020).

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Skills without a Reward Function. *CoRR* abs/1802.06070 (2018). arXiv:1802.06070

Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. Stochastic Neural Networks for Hierarchical Reinforcement Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Steven Hansen, Will Dabney, André Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. 2020. Fast Task Inference with Variational Intrinsic Successor Features. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*. 1109–1117.

Riashat Islam, Raihan Seraj, Pierre-Luc Bacon, and Doina Precup. 2019. Entropy regularization with discounted future state distribution in policy gradient methods. *arXiv preprint arXiv:1912.05104* (2019).

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research* 49, 10 (2009), 1295–1306.

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. 2019. EMI: Exploration with Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 3360–3369. `http://proceedings.mlr.press/v97/kim19a.html`

Jaekyeom Kim, Seohong Park, and Gunhee Kim. 2021. Unsupervised Skill Discovery with Bottleneck Option Learning. In *International Conference on Machine Learning*. PMLR, 5572–5582.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. 2020. Active world model learning with progress curiosity. In *International conference on machine learning*. PMLR, 5306–5315.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. 2019. Efficient Exploration via State Marginal Matching. *arXiv preprint arXiv:1906.05274* (2019).

Andrew Levy, Robert Platt, and Kate Saenko. 2019. Hierarchical Reinforcement Learning with Hindsight. In *International Conference on Learning Representations*.

Siyuan Li, Jin Zhang, Jianhao Wang, and Chongjie Zhang. 2021a. Efficient Hierarchical Exploration with Stable Subgoal Representation Learning. *arXiv preprint arXiv:2105.14750* (2021).

Siyuan Li, Lulu Zheng, Jianhao Wang, and Chongjie Zhang. 2021b. Learning Subgoal Representations with Slow Dynamics. In *International Conference on Learning Representations*. `https://openreview.net/forum?id=wxRwhSdORKG`

Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer* 21, 3 (1988), 105–117.

Daniel Ying-Jeh Little and Friedrich Tobias Sommer. 2013. Learning and exploration in action-perception loops. *Frontiers in neural circuits* 7 (2013), 37.

Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. 2017. Count-Based Exploration in Feature Space for Reinforcement Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 2471–2478. `https://doi.org/10.24963/ijcai.2017/344`

Amy McGovern and Andrew G Barto. 2001. Automatic discovery of subgoals in reinforcement learning using diverse density. (2001).

Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, and Sergey Levine. 2018. Data-Efficient Hierarchical Reinforcement Learning. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). 3303–3313.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. 2018. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*. 9209–9220.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. 2017. Count-Based Exploration with Neural Density Models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 2721–2730. `http://proceedings.mlr.press/v70/ostrovski17a.html`

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. 2007. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation* 11, 2 (2007), 265–286.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, Vol. 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. 2019. Self-Supervised Exploration via Disagreement. In *International Conference on Machine Learning*. 5062–5071.

Jean Piaget and Margaret Cook. 1952. *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. 2020. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*. PMLR, 7750–7761.

Vitchyr Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. 2020. Skew-Fit: State-Covering Self-Supervised Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 7783–7792.

Richard M Ryan and Edward L Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* 25, 1 (2000), 54–67.

Jürgen Schmidhuber. 1991. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. IEEE, 1458–1463.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. 2020. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*. PMLR, 8583–8592.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2021. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*. PMLR, 9443–9454.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. 2020. Dynamics-Aware Unsupervised Discovery of Skills. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Pranav Shyam, Wojciech Jaskowski, and Faustino Gomez. 2019. Model-Based Active Exploration. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 5779–5788. `http://proceedings.mlr.press/v97/shyam19a.html`

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. 2003. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences* 23, 3-4 (2003), 301–321.

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. 2011. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*. Springer, 41–51.

Ruo Yu Tao, Vincent François-Lavet, and Joelle Pineau. 2020. Novelty Search in Representational Space for Sample Efficient Exploration. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, and Volodymyr Mnih. 2019. Unsupervised Control Through Non-Parametric Discriminative Rewards. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. `https://openreview.net/forum?id=r1eVMnA9K7`

Yao Yao, Li Xiao, Zhicheng An, Wanpeng Zhang, and Dijun Luo. 2021. Sample Efficient Reinforcement Learning via Model-Ensemble Exploration and Exploitation. *arXiv preprint arXiv:2107.01825* (2021).

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. 2021. Reinforcement learning with prototypical representations. *arXiv preprint arXiv:2102.11271* (2021).

Chuheng Zhang, Yuanying Cai, Longbo Huang, and Jian Li. 2021. Exploration by Maximizing Renyi Entropy for Reward-Free RL Framework. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. 10859–10867.

Jesse Zhang, Haonan Yu, and Wei Xu. 2020. Hierarchical Reinforcement Learning by Discovering Intrinsic Options. In *International Conference on Learning Representations*.

Rui Zhao, Xudong Sun, and Volker Tresp. 2019. Maximum entropy-regularized multi-goal reinforcement learning. In *International Conference on Machine Learning*. PMLR, 7553–7562.

Rui Zhao and Volker Tresp. 2019. Curiosity-driven experience prioritization via density estimation. *arXiv preprint arXiv:1902.08039* (2019).

A Appendix

A.1 Surprise as prediction error

To avoid the need of the true forward model in our surprise definition, the agent can omit the left-hand term of Equation 1b by assuming the true forward model is modelled as a deterministic forward model. In this case, we can write:

$$I(S'; \Phi_T | h, A, S) \propto \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h), \phi_T \sim p(\cdot) \\ s' \sim p(\cdot|s,a,\phi_T)}} - \log p(s'|s,a,h) \tag{11a}$$

$$= \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h),\, \phi_T \sim p(\cdot) \\ s' \sim p(\cdot|s,a,\phi_T)}} - \log \sum_{\phi \in \Phi} p(s'|h,s,a,\phi) p(\phi|h) \tag{11b}$$

$$\geq \mathbb{E}_{\substack{\phi_T \sim p(\cdot),\, (s,a)\sim p(\cdot|h) \\ s' \sim p(\cdot|s,a,\phi_T), \phi \sim p(\cdot|h)}} - \log p(s'|h,s,a,\phi) \tag{11c}$$

where we applied the Jensen inequality in Equation 11c and $\phi_T \sim p(\cdot)$ is fixed. One can model $p(s'|h,s,a,\phi)$ with a unit-variance Gaussian distribution in order to obtain a tractable loss. This way, we have:

$$\mathbb{E}_{\substack{(s,a)\sim p(\cdot|h),\, \phi_T \sim p(\cdot) \\ s' \sim p(\cdot|s,a,\phi_T),\, \phi \sim p(\cdot|h)}} - \log p(s'|\phi,h,a,s) \approx \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h),\, s' \sim p(\cdot|s,a,\phi_T) \\ \phi \sim p(\cdot|h),\, \phi_T \sim p(\cdot)}} - \log \frac{1}{(2\pi)^{d/2}} e^{-0.5(s'-\hat{s}')^T (s'-\hat{s}')} \tag{12a}$$

$$\propto \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h),\, s' \sim p(\cdot|s,a,\phi_T) \\ \phi \sim p(\cdot|h),\, \phi_T \sim p(\cdot)}} ||s' - \hat{s}'||_2^2 + Const \tag{12b}$$

where

$$\hat{s}' = \arg\max_{s'' \in S} p(s''|h,a,s,\phi) \tag{13}$$

represents the mean prediction and ϕ parameterizes a deterministic forward model. Following the objective, we can extract a generic intrinsic reward as:

$$R(s,a,s') = ||f(s') - f(\hat{s}')||_2^2. \tag{14}$$

A.2 Surprise as learning progress

Let's analyze the following approximation:

$$\log p(s'|s,a,h,\phi_T) - \log p(s'|s,a,h) \approx \log p(s'|s,a,h') - \log p(s'|s,a,h). \tag{15}$$

As h' becomes large enough and the agent updates its forward model, its forward model converges to the true transition model. Formally, if one stochastic forward model can describe the transitions, we can write:

$$\lim_{|h'|\to\inf} p(s'|s,a,h') = \lim_{|h'|\to\inf} \sum_{\Phi} p(s'|s,a,h',\phi)p(\phi|h') \\ = p(s'|s,a,h',\phi_T). \tag{16a}$$

Finally, applying the procedure of Appendix A.1 in both terms of Equation 15, we end up with

$$R(s,a,s') = \Delta||s' - \hat{s}'||_2^2, \tag{17}$$

i.e, the change of prediction error caused by some interactions.

A.3 Surprise as expected information gain

Let's approximate our surprise by the expected information gain:

$$IG(h,A,S',S,\Phi) \approx IG(h,A,S',S,\Phi_T). \tag{18}$$

Then, the expected information gain Sun et al. (2011); Little and Sommer (2013) over a forward or density model parameterized by ϕ can be formulated as:

$$IG(h,A,S',S,\Phi) = I(S';\Phi|h,A,S) = \mathbb{E}_{\substack{(s,a)\sim p(\cdot|h) \\ s'\sim p(\cdot|s,a,h)}} D_{KL}(p(\Phi|h,s,a,s')||p(\Phi|h)) \tag{19a}$$

$$\approx \mathbb{E}_{\substack{(s,a)\sim\pi \\ s'\sim p(\cdot|s,a,h,\phi_T)}} D_{KL}(p(\Phi|h,s,a,s')||p(\Phi|h)) \tag{19b}$$