UPSIDE DOWN FREE ENERGY

A PREPRINT

Trenton Bricken*
Harvard University

January 2, 2022

ABSTRACT

The "Free Energy Principle" (FEP) proposes a utility function for life, which makes predictions that are surprisingly convergent with a number of recent developments in reinforcement learning (RL) [1] [2]. This paper makes three main contributions towards the application of the FEP to RL: (i) highlighting a natural connection between FEP and the recent development of "Upside Down RL" [3] [4] [5]; (ii) introducing hierarchical goals for an FEP agent to desire with a distinction between learned and evolutionary components; (iii) investigating how the theoretical lens of FEP can improve Upside Down RL beyond its current success.

***K*eywords** Free Energy Principle · Reinforcement Learning · Upside Down RL · Active Inference · Deep Learning

1 Background and Motivation

1.1 The Free Energy Principle

The Free Energy Principle (FEP) makes the key assumption that the highest reward states for an agent are those which are the least surprising (they have the highest probability of occurring according to the agent's model of the world) [6]. This assumption is justified by the observation that in order to survive, all organisms must constantly fight entropy. This entropy (or surprise) minimization is accomplished in two complementary ways: (i) building an accurate model of the surrounding world. For example, the world model of a fish should have a very strong expectation that it is always in water [1]; (ii) inferring what series of actions are necessary for the organism to minimize its entropy now and in the future. By building an accurate model of the environment and taking action to remain in the most predictable states, organisms are able to carve out environmental niches robust to the forces of nature.

[find a way to say that this encompasses predictive processing theory, bayesian brain and predictive coding.]

Note, this resistance to omnipresent entropic forces occurs at many levels. For example, the organism should model and take action to maintain its cellular integrity, blood temperature and hunger level just as much as it should find a safe place to sleep.

In addition, when presented with a novel (surprising) stimuli, there is a trade-off between the organism updating its world model to incorporate the novel information, or taking action to reject/avoid the surprising stimuli. For example, you, reader, have a very strong belief in the existence of gravity because of the vast and unconflicting evidence of gravity that has accumulated over your life. If all of a sudden a nearby object began levitating, you would have two ways to respond: You could update your belief in gravity, or take action to reject and "explain away" this surprising challenge to your beliefs. In this case, because your belief in gravity is so strong, you would first take a great number of actions to explain away your surprise, returning yourself to an unsurprising state. For example, it would be reasonable to first see if there isn't an invisible string suspending the object. However, if after taking a large number of actions the

*correspondence to trentonbricken@g.harvard.edu

phenomenon of the floating object remained unexplained, you would be forced to update your world model so that future levitations are less surprising [2].

The dual objectives of constructing an accurate world model and taking actions to satisfy it in practice produce a number of assumptions that are excitingly convergent with recent developments in model based reinforcement learning [7] [8] [9] [3] [4] [5]. A number of these have been reviewed in previous work [2].

1.2 Evolutionary priors and learned posteriors

If the FEP is the utility function for life (or a component of it) we should expect evolution to optimize for our ability to build accurate world models. This optimization would occur through our genome to provide us with evolutionary priors. For low dimensional and predictable stimuli like our body temperature, our genome is sufficient to set a very strong prior that does not need to be updated from experience. For high dimensional stimuli such as an image on our retina of a colorful mushroom, evolution can only provide us with priors in the form of inductive biases to be able to learn from experience and others what is unexpected and, by proxy, dangerous [3]. As will be outlined in Section **??**, this distinction between low dimensional evolutionary priors and high dimensional learned posteriors is one of the motivations and contributions of this paper.

1.3 Upside Down RL

Upside Down RL (UDRL), also referred to as "Reward Conditioned Policies" is a novel RL framework which makes the problem of deciding what action to take to maximize expected reward into a supervised learning problem [3] [4] [5]. The agent learns what action it should take, conditioned on its current state and a set of provided desires. For example, the agent can desire to obtain a final reward of +1,000 points, to be in a terminal state where it is resting on a platform, and to achieve this within the next 10 time steps. Learning to map these desires onto actions is distinct from traditional RL where the agent is trained to maximize its expected reward by learning to map actions onto rewards. This mapping is either explicit (through a forward planning model or Q function) or implicit (through policy gradients). This difference in approach is a figure ground inversion; one way of thinking about it is:

"akin to that between consulting a shopping list (thus letting the list determine the contents of the shopping basket) and listing some actually purchased items (thus letting the contents of the shopping basket determine the list)." [10] [11] [Cite Anscombe too]

Starting with random actions, the agent stores pairs of states, actions, and their outcomes (such as the reward received). These states and outcomes condition the policy to predict the action that produces them. By desiring close to the upper bound on what the agent has previously experienced for its next environmental rollouts, the agent iteratively improves its desires and performance.

The FEP setting where the agent knows explicitly what it desires but not what actions to take in order to obtain them lends itself naturally to UDRL. Furthermore, there is compelling pyschological evidence that our predictions about the world are in the form of sense data. What we expect to see and feel. Cite Surfing Uncertainty and SSC where talks about moving your arm and experiments around this that are meant to be in chapters 4 and 5?. This is appealing because in FEP and in the real world, we know what we desire but not how to obtain it, and must learn the series of actions to be taken.

2 FEP in Practice

2.1 Minimizing surprise by modelling the world

In this section we mathematically formalize the FEP and outline the approaches that make it tractable.

Information on what state an organism exists in is only accessible via its sensory organs. Because of observational noise and the limited capacity of these sensory organs, we assume our observations $\mathbf{o} \in O$ come from the function:

$$\mathbf{o} = g(\mathbf{s}^*; \theta) + \mathbf{z}$$

where $\mathbf{z}$ is the sensory noise, $\mathbf{s}^*$ is the true underlying state and $g()$ is the transformation of the state by sensory organs which are parameterized by θ. This noisy mapping makes the entropy of the observations an upper bound on the entropy of the true underlying states (See Appendix I for a proof of the following inequality):

[2] See https://slatestarcodex.com/2017/09/05/book-review-surfing-uncertainty/ for interesting examples where part of our actions is in fact interpreting external stimuli, for example using our world model we miss certain subtle cues.

[3] This is a real example in which infants are in fact more cautious of plants than other objects suggesting real evolutionary priors.

$$H(S^*) \leq H(O) - \int_{\mathbf{s}^* \in S^*} p(\mathbf{s}^*) \log |\frac{\partial g(\mathbf{s}^*; \theta)}{\partial \mathbf{s}^*}| \, d\mathbf{s}^* + H(S^*|O)$$

$$H(S^*) \leq H(O) - \int_{\mathbf{s}^* \in S^*} p(\mathbf{s}^*) \log |\frac{\partial g(\mathbf{s}^*; \theta)}{\partial \mathbf{s}^*}| \, d\mathbf{s}^* \tag{1}$$

This inequality exists because $H(S^*|O) \geq 0$ with equality when $S^* == O$ but this will not be the case here because of the noise $\mathbf{z}$.

The equation for the entropy of a discrete random variable is:

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)}$$

And can be thought of as the average amount of information or surprise in a random variable where $\sum_{x \in X} p(x)$ is the average (or expectation) and $\log \frac{1}{p(x)}$ is the information content of an event and the mathematical quantification of surprise.

The second term of Eq. 1:

$$\int_{\mathbf{s}^* \in S^*} p(\mathbf{s}^*) \log |\frac{\partial g(\mathbf{s}^*; \theta)}{\partial \mathbf{s}^*}| \, d\mathbf{s}^*$$

is the change of variables formula which represents the sensitivity of the organism's sensory organs to a change of the underlying state. For our purposes, we assume that evolution has refined our sensory organs to make this upper bound as tight as possible ensuring that $g(\mathbf{s}^*; \theta)$ is optimal and noise is filtered out. Therefore, we are interested in the entropy of the observations produced by our sensory organs: $H(O)$. The maximum entropy of any random variable is when it follows a uniform distribution where $p(\mathbf{o}) = 1/|O|$ such that $H(O) = |O|$ where $|O|$ is the cardinality of O which creates an upper bound on the entropy $H(O) \leq |D|$ (See Appendix II for a proof). Thus, to minimize $H(O)$, the frequency of observations must be maximally concentrated on the smallest subset of observations. In other words, using the mathematical quantification of surprise $\log \frac{1}{p(\mathbf{o})}$, by minimizing $H(O)$ we are minimizing the average surprise of our observations and the upper bound entropy of our underlying states.

However, in order to minimize $H(O)$ and satisfy the FEP, we must be able to compute $p(\mathbf{o})$. $p(\mathbf{o})$ represents many different forms of sensory observation and is very high dimensional making it intractable to compute explicitly. Instead, it can be computed using the latent variable $\mathbf{s}$ which can be thought of as the internal representation the agent has of the true underlying state $\mathbf{s}^*$ such that:

$$p(\mathbf{o}) = \int_{\mathbf{s} \in S} p(\mathbf{o}|\mathbf{s}) p(\mathbf{s}) \, d\mathbf{s} \tag{2}$$

Use of this latent variable is not only helpful for computing $p(\mathbf{o})$ but also theoretically supported by the fact the brain is a generative model as evidenced by our imagination. However, Eq. (2) is also intractable because $\mathbf{s}$ is very often high dimensional. As a result, Friston proposes using variational inference to maximize a lower bound on $p(\mathbf{o})$ which corresponds to minimizing an upper bound on surprise or entropy. The intuition behind using variational inference is that rather than needing to compute the integral over every possible $\mathbf{s}$, to only compute the integral over those $\mathbf{s}$ that have the highest probability of occurring $p(\mathbf{s}|\mathbf{o})$ and thus obtaining a lower bound:

$$p(\mathbf{o}) \geq \mathbb{E}_{\mathbf{s} \in p(\mathbf{s}|\mathbf{o})}[p(\mathbf{o}|\mathbf{s})]$$

Because, $p(\mathbf{s}|\mathbf{o})$ is intractable ($p(\mathbf{s}|\mathbf{o}) = \frac{p(\mathbf{s}, \mathbf{o})}{p(\mathbf{o})}$ and we would need to know $p(\mathbf{o})$, the original problem!) we approximate it by minimizing the Kullback-Leibler divergence between $p(\mathbf{s}|\mathbf{o})$ and a latent probability distribution to be learned $q(\mathbf{s}; \theta)$ with the sufficient statistics θ. For this minimization to be tractable we restrict $q(\mathbf{s}; \theta)$ to a given model class, in most cases Gaussian distributions because it is an expressive model class that is efficient to work with, for example it has a closed form solution to the Kullback-Leibler divergence [12]. This minimization and its result in lower bounding $p(\mathbf{o})$ is as follows (See Appendix III for an alternative derivation that utilizes Jensen's inequality.):

$$D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})] = \int_{\mathbf{s} \in S} q(\mathbf{s}) \log \frac{q(\mathbf{s})}{p(\mathbf{s}|\mathbf{o})} ds$$

$$= \int_{\mathbf{s}\in S} q(\mathbf{s}) \log q(\mathbf{s}) ds - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{s}|\mathbf{o}) ds$$

$$= \int_{\mathbf{s}\in S} q(\mathbf{s}) \log q(\mathbf{s}) ds - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log \frac{p(\mathbf{o}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{o})} ds$$

$$= \int_{\mathbf{s}\in S} q(\mathbf{s}) \log q(\mathbf{s}) ds - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{o}|\mathbf{s}) ds - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{s}) ds + \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{o}) ds$$

$$= D_{KL}[q(\mathbf{s})||p(\mathbf{s})] - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{o}|\mathbf{s}) ds + \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{o}) ds$$

$$D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})] = D_{KL}[q(\mathbf{s})||p(\mathbf{s})] - \int_{\mathbf{s}\in S} q(\mathbf{s}) \log p(\mathbf{o}|\mathbf{s}) ds + \log p(\mathbf{o})$$

$$D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})] = D_{KL}[q(\mathbf{s})||p(\mathbf{s})] - \mathbb{E}_{\mathbf{s}\in q(\mathbf{s})}[\log p(\mathbf{o}|\mathbf{s})] + \log p(\mathbf{o})$$

$$\log p(\mathbf{o}) - D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})] = \mathbb{E}_{\mathbf{s}\in q(\mathbf{s})}[\log p(\mathbf{o}|\mathbf{s})] - D_{KL}[q(\mathbf{s})||p(\mathbf{s})]$$

$$\log p(\mathbf{o}) \geq \mathbb{E}_{\mathbf{s}\in q(\mathbf{s})}[\log p(\mathbf{o}|\mathbf{s})] - D_{KL}[q(\mathbf{s})||p(\mathbf{s})]$$

With the last step because KL Divergence is always positive: $D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})] \geq 0$

For this minimization to work we need to sample observations from our dataset D of observations from agent interactions with the environment, and compute expectations $\mathbb{E}_{\mathbf{o}\in D}$ resulting in the objective:

$$\mathbb{E}_{\mathbf{o}\in D}[\log p(\mathbf{o}) - D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{o})]] = \mathbb{E}_{\mathbf{o}\in D}[\mathbb{E}_{\mathbf{s}\in q(\mathbf{s})}[\log p(\mathbf{o}|\mathbf{s})] - D_{KL}[q(\mathbf{s})||p(\mathbf{s})]]$$

To make $q(\mathbf{s})$ closer to $p(\mathbf{s}|\mathbf{o})$ it makes sense to also condition $q(\mathbf{s})$ on the available observation $\mathbf{o}$ making $q(\mathbf{s})$ into the conditional probability $q(\mathbf{s}|\mathbf{o})$.

Bringing everything together, we want to compute:

$$p(\mathbf{o}) = \int_{\mathbf{s}\in S} p(\mathbf{o}|\mathbf{s}) p(\mathbf{s})\, d\mathbf{s}$$

The way to do this tractably is to only integrate over the highest probability states $p(\mathbf{s}|\mathbf{o})$ which we approximate with $q(\mathbf{s}|\mathbf{o})$ by minimizing the Kullback-Leibler divergence, $D_{KL}[q(\mathbf{s}|\mathbf{o};\theta)||p(\mathbf{s}|\mathbf{o})]$. By performing this minimization, taking gradients with respect to the sufficient statistics θ of the Gaussian distribution q(s), we minimize the objective function:

$$\mathbb{E}_{\mathbf{o}\in D}[\mathbb{E}_{\mathbf{s}\in q(\mathbf{s}|o;\theta)}[\log p(\mathbf{o}|\mathbf{s})] - D_{KL}[q(\mathbf{s}|o;\theta)||p(\mathbf{s})]]$$

And by doing so we obtain a lower bound on $p(\mathbf{o})$ which becomes tighter the smaller $D_{KL}[q(\mathbf{s}|o;\theta)||p(\mathbf{s}|\mathbf{o})]$ is.

When this lower bound is made as tight as it can be we compute an unbiased estimate of $p(\mathbf{o})$ using Importance Sampling:

$$p(\mathbf{o}) = \int_{\mathbf{s}\in S} p(\mathbf{o}|\mathbf{s}) p(\mathbf{s}) \frac{p(\mathbf{s}|\mathbf{o})}{p(\mathbf{s}|\mathbf{o})}\, d\mathbf{s}$$

$$= \mathbb{E}_{\mathbf{s}\sim p(\mathbf{s}|\mathbf{o})}[\frac{p(\mathbf{o}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{s}|\mathbf{o})}]$$

2.2 Minimizing surprise by taking action

Learning an accurate model of the world is insufficient to minimize entropy in a stochastic and naturally entropic environment. An organism can be forcefully displaced by the weather, need for food, or an approaching lion. Moreover, certain actions such as the construction of shelter, while creating a new and surprising environment, can significantly reduce entropy in the long run [13].

According to the FEP, with a world model present, the organism uses "Active Inference" to take actions that minimize its present and expected future surprise. How the agent should minimize its expected future surprise is an area of active research with a number of interesting proposals [1] [14] [13] [15] [16] [17]. In this paper we take the same approach as [13] by learning a policy π that seeks to satisfy the FEP directly, learning to take actions that least surprise the learned world model. This approach lacks incorporation of an exploratory term that other implementations have and we leave this for an area of future work.

Our policy is $\pi(o;\phi)$ where $\mathbf{o}$ is the current state of the agent and ψ are the parameters of the function (in this case a neural network) that outputs what action should next be taken. And we seek to minimize w.r.t ϕ the surprise of our observations across an agent's whole lifetime which we denote with subscripts $t = 0$ and $T = terminal$:

$$\arg\min_{\phi} -\log p(\mathbf{o}_{t:T}|\pi) = -\sum_{t}^{T-1} \log p(\mathbf{o}_{t+1}|\mathbf{o}_{\leq t}, \pi)$$

We make the Markov Assumption:

$$-\log p(\mathbf{o}_{t:T}|\pi) = -\sum_{t}^{T-1} \log p(\mathbf{o}_{t+1}|\mathbf{o}_{t}, \pi)$$

In the setting of UDRL, this optimization is straightforward because we train our policy to produce the action that corresponds to our specific desires which are what would minimize surprise.

2.3 Evolutionary Bootstrapping

We have used bold notation to denote that the sensory observations $\mathbf{o}$ are a vector that contains stimuli from different channels. For the state, non-pixel based environments these stimuli consist of the bodily state denoted v and sensory rewards r. For the pixel based environment, we use the visual observations instead of bodily state to represent v. In principle, these two sensory channels are both learned over time using the Free Energy principle and should be treated in a Bayesian way:

$$p(\theta|\mathbf{o}_t) = \frac{p(\mathbf{o}_t|\theta)p(\theta)}{p(\mathbf{o}_t)} \tag{3}$$

Where $p(\theta)$ is the prior (before any posterior updates this is the "evolutionary information"). $p(\mathbf{o}_t|\theta)$ is the likelihood of the observation under the current model parameters and $p(\theta|\mathbf{o}_t)$ are the posterior model parameters after updating to the current observation from time t. This posterior will act as the new prior for future observations.

While both the visual and reward channels should be updated in this way, in reality they are both represented quite differently. The reward channel can be thought of, without loss of generality, as a scalar, real valued sensation, like hunger. Humans (and likely all living organisms) have evolved over the eons to have a very strong prior expectation that they will not be hungry. A lack of food in one's belly is therefore surprising and our prior for not being hungry is so strong that no amount of Bayesian updating will shift our prior such that we are no longer surprised and displeased by the onset of hunger[4]. Because the reward channel is low dimensional and, for the environments used here a scalar, it is easy for evolution to learn a prior probability distribution. We refer to this prior as "evolutionary bootstrapping" because in the long run, we expect that an agent without a strong prior for not being hungry would learn over time by applying the FEP, that a lack of food corresponds to surprising states of immobility and death and should thus be avoided. Therefore, we assume that learning a model with FEP would converge and result in the same solution that our evolutionary prior already provides us with. As a result, we can make learning more efficient and boostrap our agent by

[4]There are accounts of people who are able to learn to ignore the sensations of hunger. However, this is not only incredibly difficult to do but requires overriding the effect rather than changing one's evolutionary ingrained biological signalling and displeasure of hunger.

providing this evolutionary prior before any training has occurred. Providing this prior is also useful for the many RL environments which are not naturally entropic, in order to motivate the desired actions, for example, the car racing environment used in [7] and the mountain car environment from [14].

Meanwhile, the visual sensory channel, v is too high dimensional for evolution to reliably bootstrap a useful prior[5]. For example, there are a plethora of different foods that all look different and can be eaten in different environments to satisfy hunger equally well. As a result, we assume that the prior for v is uniform or normally distributed for the sake of regularization. As a result, it is the learned likelihood $p(v|\theta)$ acquired from experience in the environment that determines which states are seen as less surprising and desirable.

For these reasons, we ignore the Bayesian updating in Eq. 3 by assuming that $p(\theta_r|r) = p(\theta_r)$ and $p(\theta_v|v) = p(v|\theta_v)$. Given that we thus have our model of $p(r)$ we write $p(\mathbf{o}) = p(v, r) = p(v|r)p(r)$ and learn $p(v|r)$ as a conditional VAE. By modelling the relation between v and r, which certainly exists, rather than assuming they are independent, we will be able to learn what visual stimuli correspond to various sensory reward states. For example, using our hunger example again, conditioning upon a state of satiation, we could generate images of satiating foods like cake or pizza.

3 Loss Function Interpretations

What other interpretations can I incorporate?

Different paper on FEEF and exploration policies and how they are part of FEP.

Looking at the FEEF objective function in different ways:

$$KL[p(o_{t:T}|\pi)||\tilde{p}(o_{t:T})] = \mathbb{E}_{p(o_{t:T}|\pi)}[\log p(o_{t:T}|\pi) - \log \tilde{p}(o_{t:T})]$$

$$= -\underbrace{H(p(o_{t:T}|\pi))}_{\#1} - \underbrace{\mathbb{E}_{p(o_{t:T}|\pi)}[\log \tilde{p}(o_{t:T})]}_{\#2} \tag{4}$$

#1 Seeks to maximize entropy and by proxy the average surprise and information gain under the world model from actions taken by the policy.

#2 Takes actions to satisfy the biased model's desires by taking actions that are not surprising under this biased model.

We can re-write $\tilde{p}(o_{t:T})$ applying our Bayesian updating rule in Eq. 3 but with Bayes rule in a different form as:

$$\tilde{p}(o_{t:T}) = \mathbb{E}_{p(\theta)}[\tilde{p}(o_{t:T}|\theta)] = \mathbb{E}_{p(\theta)}[\frac{p(\theta|o_{t:T})p(o_{t:T})}{p(\theta)}]$$

Plugging this into Eq. 4:

$$= -H(p(o_{t:T}|\pi)) - \mathbb{E}_{p(o_{t:T}|\pi)}[\log \mathbb{E}_{p(\theta)}[\frac{p(\theta|o_{t:T})p(o_{t:T})}{p(\theta)}]]$$

Using Jensen's inequality for the concave natural logarithm: $\mathbb{E}[\log(x)] \leq \log(\mathbb{E}[x]))$

$$\log \mathbb{E}_{p(\theta)}[\frac{p(\theta|o_{t:T})p(o_{t:T})}{p(\theta)}] \geq \mathbb{E}_{p(\theta)}[\log p(\theta|o_{t:T}) + \log p(o_{t:T}) - \log p(\theta)]$$

Given that we are trying to minimize $KL[p(o_{t:T}|\pi)||\tilde{p}(o_{t:T})]$ w.r.t π we want $-\mathbb{E}_{p(o_{t:T}|\pi)}[\log \mathbb{E}_{p(\theta)}[\frac{p(\theta|o_{t:T})p(o_{t:T})}{p(\theta)}]]$ to be as large as possible and so our Jensen's inequality giving a lower bound will correctly force the equation we are approximating to be a maximum.

$$= -H(p(o_{t:T}|\pi)) - \mathbb{E}_{p(o_{t:T}|\pi)}[\mathbb{E}_{p(\theta)}[\log p(\theta|o_{t:T}) + \log p(o_{t:T}) - \log p(\theta)]]$$

[5]There are some exceptions which prove the rule. For example, baby birds immediately upon birth are able to recognize the wing patterns of predatory birds and hide. However, they can also be fooled by cardboard cutouts of these bird's silhouettes.

$$= -H(p(o_{t:T}|\pi)) - \underbrace{\mathbb{E}_{p(o_{t:T}|\pi),p(\theta)}[\log p(\theta|o_{t:T})]}_{\#1} - \underbrace{\mathbb{E}_{p(o_{t:T}|\pi)}[\log p(o_{t:T})]}_{\#2} - \underbrace{H(p(\theta))}_{\#3}$$

#1 Seeks to minimize the entropy of the posterior desires, therefore becoming confident in what is desired.

#2 Actions should lead to observations that satisfy the world model (VAE) such that observations are not surprising.

#3 $\mathbb{E}_{p(\theta)}[\log p(\theta)] = -H(p(\theta))$. Maximize entropy of the prior desires. NB. this term will be dropped when optimizing for the policy π.

We can look at this equation an additional way by noting that the first and third terms:

$$-H(p(o_{t:T}|\pi)) - \mathbb{E}_{p(o_{t:T}|\pi)}[\log p(o_{t:T})] = KL[p(o_{t:T}|\pi)||p(o_{t:T})]$$

And the second and 4th terms:

$$-\mathbb{E}_{p(o_{t:T}|\pi),p(\theta)}[\log p(\theta|o_{t:T})] - H(p(\theta)) = \mathbb{E}_{p(o_{t:T}|\pi)}[KL[p(\theta)||p(\theta|o_{t:T})]]$$

Bringing everything back together we have:

$$KL[p(o_{t:T}|\pi)||\tilde{p}(o_{t:T})] = \underbrace{KL[p(o_{t:T}|\pi)||p(o_{t:T})]}_{\#1} + \underbrace{\mathbb{E}_{p(o_{t:T}|\pi)}[KL[p(\theta)||p(\theta|o_{t:T})]]}_{\#2}$$

#1 Policy seeks to minimize the difference between the observations it generates and those expected by the world model.

#2 Policy seeks to minimize the difference between the prior and posterior desires. In other words, the generated observations should fit the prior desires (note that this prior is the previous posterior before time point t).

4 Future Work

Also compelling because of mimickry. We aren't smarter just better at learning from others and mimicking them. Can learn given our desires, what actions to take in order to obtain them from others. Know what you want and take specific actions at specific time points. Also mimickry in humans. Mimick and learn from the actions of previous rollouts.

[would like for there to be a section on explore vs exploit tradeoffs.

This is one way in which the Free Energy Principle, by coining everything in a Bayesian, information theoretic framework, it is able to balance exploration and exploitation. (See bacteria chemoattractant paper for further examples).
]

5 Conclusion

Acknowledgements

I would like to thank Beren Millidge and Alec Tschantz for providing lots of inspiration and advice during this project. I would also like to thank open source software contributors, including but not limited to: Numpy, Pandas, Scipy, Matplotlib, PyTorch, and Anaconda.

References

[1] Karl J. Friston, Jean Daunizeau, and Stefan J. Kiebel. Reinforcement Learning or Active Inference? *PLoS ONE*, 4(7):e6421, July 2009.

[2] Beren Millidge. Deep Active Inference as Variational Policy Gradients. *arXiv:1907.03876 [cs]*, July 2019. arXiv: 1907.03876.

[3] Rupesh Kumar Srivastava, Pranav Shyam, Filipe Mutz, Wojciech Jaśkowski, and Jürgen Schmidhuber. Training Agents using Upside-Down Reinforcement Learning. *arXiv:1912.02877 [cs]*, December 2019. arXiv: 1912.02877.

[4] Juergen Schmidhuber. Reinforcement Learning Upside Down: Don't Predict Rewards – Just Map Them to Actions. *arXiv:1912.02875 [cs]*, June 2020. arXiv: 1912.02875.

[5] Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-Conditioned Policies. *arXiv:1912.13465 [cs, stat]*, December 2019. arXiv: 1912.13465.

[6] Samuel J Gershman. What does the free energy principle tell us about the brain? page 10.

[7] David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122 [cs, stat]*, March 2018. arXiv: 1803.10122.

[8] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *CoRR*, abs/1811.04551, 2018.

[9] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv:1912.01603 [cs]*, March 2020. arXiv: 1912.01603.

[10] Andy Clark. *Surfing uncertainty: prediction, action, and the embodied mind*. Oxford University Press, 2019.

[11] Scott Alexander. book review: surfing uncertainty | slate star codex.

[12] David M. Blei, Alp Kucukelbir, and Jon D. Mcauliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[13] Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. SMiRL: Surprise Minimizing RL in Dynamic Environments. *arXiv:1912.05510 [cs, stat]*, February 2020. arXiv: 1912.05510.

[14] Kai Ueltzhöffer. Deep active inference. *Biological Cybernetics*, 112(6):547–573, December 2018.

[15] Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Learning action-oriented models through active inference. *PLOS Computational Biology*, 16(4):e1007805, April 2020.

[16] Alexander Tschantz, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. Reinforcement Learning through Active Inference. *arXiv:2002.12636 [cs, eess, math, stat]*, February 2020. arXiv: 2002.12636.

[17] Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference. *arXiv:1911.10601 [cs, eess, math, stat]*, November 2019. arXiv: 1911.10601.