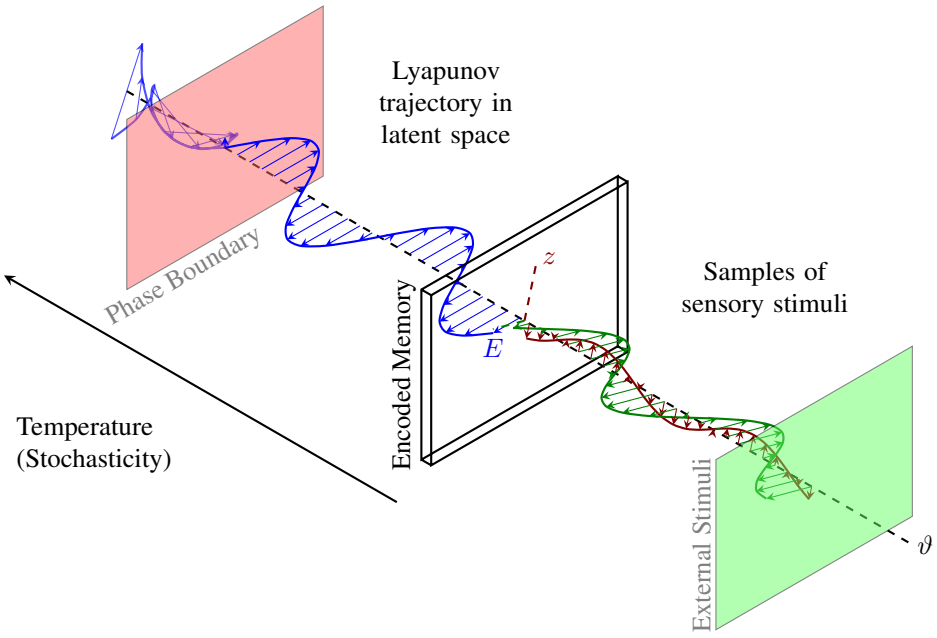
Energy-Based Dynamics in Consciousness

Luke J. Pereira

Abstract

The *free energy principle* shows us that biological dissapative systems learn to endure by acting on the environment to resist phase transitions that would otherwise change their physical structure. It is possible to create an adaptive artificial agent that reproduces this behaviour in a latent phase space that is bounded by two energy manifolds. The artificial agent can then learn how to maintain a dynamic trajectory in the phase space that minimizes the free energy of the joint energy functions represented as a Lyapunov function. This can be also expressed as minimizing the Kullback–Leibler divergence or relative entropy of the joint probability densities. The lower and upper bounds on the latent space can be created by repeating a process of learning a world model: initially on external sensory data and again on latent data, which results in a stochastic mixture due to a loss of information. A comparison can be made between the agent's dynamic trajectory in the latent space and the creation of stochastic environments in our own dreams. Our dreams, which can be described as sensorimotor hallucinatory experiences that follow narrative structures, also appear to result from a closed-loop of our action and perception mechanisms from external sensory stimuli.



A depiction of the high-dimensional energy manifolds that act as phase boundaries between the states of consciousness.

1 Overview

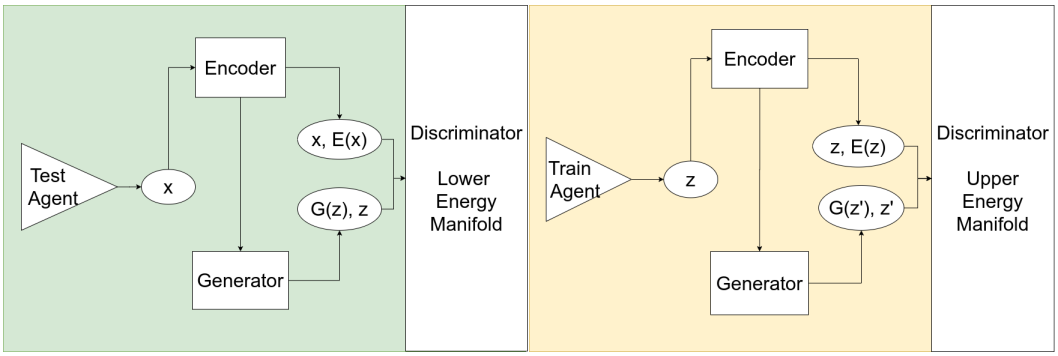
Energy-based models (EBMs) provide an alternative perspective to the standard optimization approach of using a cost function to measure a model's ability to learn a probability density. This provides a useful layer of abstraction to build on and may also be more computationally efficient by avoiding computing an intractable posterior and only computing variable dependencies. A phase space that is divided by two manifolds described by separate energy functions that represent phase boundaries of the three states in which the agent can exist is proposed. The approach of maintaining two energy functions is in contrast to the standard approach of minimizing a single cost function. Having a mixture between an true world model and an upper bound of incoherent reality allows for better predictions within a stochastic test environment and also encourages creative experimentation and imaginative exploration.

In thermodynamics, a dissipative system is an open system that exchanges energy with an environment. One notion of a dissipative system is the existence of a Lyapunov function. By dynamically optimizing the internal parameters of an agent we can minimize the free energy represented by a Lyapunov function that contains a Kullback–Leibler divergence or relative entropy of the joint energies. At higher states, the temperature or stochastic nature of the environment grows, which decreases the stability of the dissapative agent and increases the likelihood that it crosses a phase boundary and transitions its state. The degree of randomness in the environment corresponds to information loss that results from repeatedly encoding latent observations.

If we label the three states an agent can exist in as *waking* consciousness, *dreaming* consciousness, and *lucid dreaming* consciousness we establish a model of the mind. The lowest phase space is observable reality and the phase space between the two manifolds is the dream world. It is sufficiently distanced so that prediction errors, which can be interpreted as confabulation, do not necessarily push the agent into a phase transition. The lower phase boundary, dividing waking and dreaming consciousness, is the encoded discriminator's energy function, E_{world}. This can be interpreted as a collection of observations of reality from the training data and any highly plausible generated predictions. This manifold serves as the lower bound or ground truth of the agent's world model. The higher phase boundary, dividing dreaming and lucid dreaming consciousness, is the energy manifold of a highly stochastic generative model, E_{dream}. After passing the phase boundary into dream lucidity, information sampled by the agent is interpreted with access to both discriminators. In such a state, an agent is aware of both the dream world implausibility and the constraints of reality. From this vantage point, a system can temporarily distill knowledge and explore abstractions of a highly stochastic reality. By considering this metaphysical structure as a model of our consciousness, we open two paths of exploration in dream research and as an approach to developing human-like creative exploration in artificial agents.

2 Training the Energy Functions

In a standard training algorithm, we aim to minimize a loss from a cost function. In terms of our energy-based model, this would be equivalent to minimizing the space between the generative world model to make it as close to the training data as possible. This results in an artificial intelligence agent that exists as closely in the waking state as possible but lacks imaginative abilities. Instead, we train an artificial agent to have both a realistic model of the world but also be able to have an understanding of how to creatively modify its world model to some extreme. Moreover, the agent should be able to validate these modifications, store its ideas in memory, and later decode and re-enact the useful ideas on the external world environment.



There are two phases of training, waking and dreaming, which occur sequentially. In the waking phase, an agent explores the external environment and creates an accurate model of the world. It does this by first encoding its sensory data, then producing synthetic predictions using the latent representation and training a discriminator on the tuple of outputs. This discriminator will be the lower energy manifold that acts as a phase boundary between the waking and dreaming state. In the dreaming phase, the agent no longer has access to external sensory stimuli so it explores a latent environment. It performs a secondary encoding of its latent states, then generates synthetic predictions and trains a discriminator on the tuple of outputs. The dream model is able to produce highly implausible and incoherent data up to some coherency threshold. This threshold can be seen to be proportional to the dimensional difference between the information bottleneck of the encoder and the original dimension of the input. The energy function of this secondary discriminator will serve as the upper manifold, beyond which the environment becomes incoherent.

Both phases use a Bidirectional Generative Adversarial Networks (BiGAN) (Donahue et al. 2017) architecture. In addition to a generator G from the standard GAN framework (Goodfellow et al., 2014), a BiGAN includes an encoder E which maps input x to its latent representations z. The BiGAN discriminator D will jointly discriminate in both data space and latent space using tuples $(x, E(x))$ versus $(G(z), z)$. It can be proven that the BiGAN's encoder and generator must learn to invert one another in order to fool the BiGAN discriminator in an adversarial game. The process is repeated in latent space during dreaming phase with latent data being further distilled. By reusing the encoder and generator in both phases, we are able to have encoding and generative mechanisms that perform well at both levels of abstractions.

3 The Free Energy Principle

The free energy principle can be used to describe how thermodynamically open biological systems negotiate a changing or non-stationary environment in a way that allows them to endure over substantial periods of time (Friston, 2006). Let ϑ parameterize environmental forces or fields that act upon the agent and λ be quantities or temperature that describe the agents physical state. The free energy is a scalar function of the ensemble density and the current sensory input. Let $q(\vartheta; \lambda)$ be an arbitrary density function on the environments parameters that is specified or encoded by the agents parameters. It can be regarded as the probability density that a specific environmental state ϑ would be selected from an infinite ensemble of environments given the agents state λ, which is fixed and known. Then the free energy of the agent is given by,

$$\begin{aligned} F &= \int q(\vartheta) \ln \frac{p(\tilde{y}, \vartheta)}{q(\vartheta)} d\vartheta \\ &= -\langle \ln p(\tilde{y}, \vartheta) \rangle_q + \langle \ln q(\vartheta) \rangle_q \end{aligned} \tag{1}$$

Here $\langle \cdot \rangle_q$ means the expectation under the ensemble density q.

4 Training and Testing the Agent

A Lyapunov function is constructed to represent the composition of the two energy functions. A Lyapunov function is a scalar function of a systems state that decreases with time. Instead of trying to infer the Lyapunov function given an agent's structure and behaviour, we train the agent to minimize its Lyapunov function (its free energy) by optimizing its parameters. The free energy principle states that all the quantities that that are owned by the system will change to minimize free energy. These quantities are the agent's internal parameters λ and the action parameters α. We can rearrange (1) to show the dependence of the free energy on α and λ,

$$\begin{aligned} F &= -\ln p(\tilde{y}) + D(q(\vartheta;\lambda)||p(\vartheta|\tilde{y})) \\ &= -\langle \ln p(\tilde{y},\vartheta)\rangle_q + D(q(\vartheta;\lambda)||p(\vartheta|\tilde{y})) \end{aligned} \tag{2}$$

Where D is the Kullback–Leibler cross-entropy or divergence term that measures the difference between the ensemble density and the conditional density of the causes. Changing the configuration of the system to move or resample the environment by optimizing its actions α will minimize the free energy of the first term. The Kullback–Leibler divergence D is used to descend the Lyapunov free energy by optimizing the agents internal parameters λ in the second term with,

$$D(q(\vartheta;\lambda) = \int q \ln \frac{q}{p} d\vartheta.$$

By minimizing its free energy, the agent learns how to act in order for its trajectories to stabilize by descending into basins of attractions to find equilibrium points. It follows that the latent trajectories can be interpreted with symbolic dynamics. A symbolic orbit is a sequence of symbols corresponding to the successive partition elements visited by the point in its orbit. Instead, we use an Autoencoder to learn an encoding that will partition the space, which reduces the memory and computation of what would otherwise involve solving high dimensional differential equations on a continuous domain.

5 Discussion

It is conceivable that the state of a dreaming artificial intelligence agent can be pushed to transition phases into lucidity, which is a higher analog of the shallow consciousness experienced when being trained solely on sensory inputs. It's unknown what role dream lucidity plays in the development of consciousness or creativity. In a simulation game, we may attempt to use lucidity of a computer to unveil tactics or strategies that we were unaware of in order to improve our performance and our understanding of the game.

It is also conceivable that the dissapative state of lucid dreaming in humans can be extended using an external system to provide perturbations that drive a dreaming consciousness to transition phases into lucid dreaming consciousness and stabilize in this state. In this state we would be able to freely modify our world without being constrained by physical reality or our physical abilities. From here, it may be possible to store and later decode our imagined actions and ideas from dreams to reality.

References

[1] Friston, K.J., Stephan, K.E. Free-energy and the brain. Synthese 159, 417–458 (2007).

[2] J. Donahue, P. Krahenbuhl, and T. Darrell. Adversarial feature learning. arXiv preprint arXiv:1605.09782, 2016.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[4] Nir Y, Tononi G. Dreaming and the brain: from phenomenology to neurophysiology. Trends Cogn Sci. 2010;14(2):88-100. doi:10.1016/j.tics.2009.12.001