Unifying Inference in Ergodic World Models

Luke J. Pereira

Abstract

Fast and slow inference are presented as trajectories existing on a high-dimensional spectrum with their extrema bounding surfaces of a latent phase space. On one end, an Autoencoder is trained on sensory observations and impresses latent encodings onto a Discriminator-like energy manifold. On the other end, the latent space is delimited by repeating the training process on sparse and low-dimensional symbolic abstractions, producing an upper bounding manifold. An agent's trajectory corresponds to a flow of a symbolic dynamical system that minimizes the KL divergence between the upper and lower boundaries, which minimizes the free-energy of the system and can be learned from a series of initial value random walks.

Birkhoff's theorem states that in an ergodic domain, the average of a function f over an infinite number of flows converges to the continuous integral of the function over the phase space. Defining f to be a function that minimizes the joint divergence of the boundaries of a world model will produce a measure preserving ergodic flow. Moreover, we can show that the time average can be derived from inference using standard Bayesian expected loss while the space average can be derived from energy-based inference methods. This reveals that the two methods of inference become pointwise unified in ergodic domains, like language generation. This in turn explains the effectiveness of modern deep learning methods that, despite being restricted to surface level System 1 thinking, are able to imitate System 2 thinking. It is possible for a model to learn how to construct and maintain a minimal ergodic latent space even when sampling from sparse non-ergodic observations. This pointwise equivalence enables an agent to fluidly switch between systems of thinking and improve inference performance.

1 Systems of Thinking

In *Thinking, Fast and Slow* (Kahneman, 2011), cognitive processes are categorized into two modes. In System 1, thought processes are fast, intuitive, unconscious, and habitual. In System 2, thought processes become slower, logical, conscious, algorithmic, and may involve planning or reasoning. It can be claimed that the system of fast thought closely corresponds to many state-of-the-art deep learning approaches (Bengio, 2019). These methods aim to discover underlying distributions of data in the form of a normalized probability density by training on many examples while minimizing a cost function. In doing so, a model is able to quickly perform inference when being tested on unseen examples. System 2 thinking instead embodies the future aspirations of models that require very few examples and can learn in a self-supervised manner. Though the model may perform inference slowly, this method of thinking appears to be a necessary aspect in developing an artificial general intelligence (AGI) that matches and exceeds our own abilities. It's possible to draw a comparison between System 2 thinking and energy-based models (EBM) which aim to capture dependencies between variables by associating a scalar energy to each configuration of the variables (LeCunn, 2006). Learning is often faster than with a standard loss functions and consists in finding an energy function where observed configurations of the variables are given lower energies than unobserved ones. Inference is often slower and involves searching the energy function using optimization methods like stochastic gradient descent to find compatible variables that minimize the energy function.

2 Birkhoff's Ergodic Theorem

A dynamical system, usually written as the tuple (T, X), is described by a transformation that maps a phase space onto itself, $T : X \to X$. The set of points attained from repeated applications of the transformation from some starting point is known as its forward orbit or trajectory. ergodicity can be viewed as an indecomposability condition and is concerned with how a typical orbit of a dynamical system is distributed throughout the phase space with these qualitative distributional properties being expressed in terms of measure theory. Measure preserving means that $P(T^{-1}(A)) = P(A)$ for all measurable sets $A \in \mathcal{A}$. ergodic means that $T(A) = A$ implies $P(A) = 0$ or $P(A) = 1$ for all $A \in \mathcal{A}$.

Birkhoff's Ergodic Theorem states that if a mapping is ergodic, as the number of finite averages taken along any of its orbits increases to infinity (the time average), this value will converge to the continuous integral (the space average). That is, a finite average sampling of points of any orbit will be as accurate as a continuous average integral over the entire state space. Formally, let $(X, \mathcal{B}, \mu)$ be a probability space and let $T : X \mapsto X$ be a measure-preserving transformation (on a σ-finite measure space $\mu(X) < \infty$). If f is any integrable function and T is ergodic, then the time average is constant in measure μ almost everywhere and is equivalent to the space average,

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))}_{\text{Time Average, } \hat{f}} = \underbrace{\frac{1}{\mu(x)} \int_X f d\mu}_{\text{Space Average, } \bar{f}}$$

Corollary (Point-wise Ergodic Theorem):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = E(f)$$

Working backwards from the expected value, we can derive the standard Bayesian training method of averaging over a large number of training examples. We consider $x \in X$ to be a sampled observation, f is our cost minimization function, and T represents a transformation attained from a trained model that encodes and predicts a future state. As the number of training orbits increases to infinity, the encoding and prediction mechanisms converge to the continuous integration over the space and the expected output of the function in the point-wise interpretation.

Energies can be thought of as being *unnormalized negative log probabilities*. That is, we may use the Gibbs-Boltzmann distribution to convert an energy function to its equivalent probabilistic representation after normalization, i.e. $P(y \mid x)$. Recall, *marginalisation* is a method that sums over the possible values of one variable to determine the marginal contribution of another. $P(y \mid x)$ is just an application of the Gibbs-Boltzmann formula with latent variables z being marginalized implicitly through integration, i.e. $P(y \mid x) = \int_z P(y, z|x)$. Then,

$$P(y \mid x) = \frac{\int_z \exp(-\beta E(x, y, z))}{\int_y \int_z \exp(-\beta E(x, y, z))}$$

The derivation introduces a β term which is the inverse of temperature T, so as $\beta \to \infty$ the temperature goes to zero. β is a positive constant that needs to be calibrated to fit the model. A larger β value produces a more fluctuate model while a smaller β gives a smoother model. When $\beta \to \infty$, we see that $\breve{y} = \text{argmin}_y E(x, y)$. So we can redefine our energy function as an equivalent function using F_β,

$$F_\infty(x, y) = \text{argmin}_z E(x, y, z)$$
$$F_\beta(x, y) = -\frac{1}{\beta} \log \int_z \exp(-\beta E(x, y, z)).$$

In physics, F_β is known as the free energy and E is the energy. If we have a latent variable model and want to eliminate the latent variable z in a probabilistically correct way, we just need to redefine the energy function in terms of F_β,

$$P(y \mid x) = \frac{\exp(-\beta F_\beta(x, y, z))}{\int_y \exp(-\beta F_\beta(x, y, z))}.$$

3 Ergodic World Models

We see that in ergodic domains, an averaging of an increasingly large number of examples will become increasingly close to the average behaviour of the underlying causative process depicted by the space average integral. We can claim an equivalence between the time averages of a summation with the first system of thinking and the space average of an integral with the second system of thinking. It can then be claimed that the dichotomy of the systems of thought become increasingly unified in ergodic domains.

In environments that contain non-ergodic domains, a learning agent can establish an internal ergodic world model by actively selecting which observations are sampled and manipulating how these observations are encoded and stored. To maintain ergodic averages, only a local stablility is necessary, making this model biologically plausible. An ergodic world model allows the agent to attain a pointwise fluidity between System 1 and System 2 inference capabilities which ultimately improves its decision-making speed and quality in unpredictable environments. When applied to the energy-based dynamical system described in a previous paper, the latent space explored during the dream phase of training can be manipulated in order to generate ergodic measure preserving flows despite the external world explored during its waking test phase not necessarily being ergodic. Recall, the flows are attained by minimizing the free-energy of the system, represented as a Kullback–Leibler divergence or relative entropy of bounding energy manifolds.

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 is typically be represented as a Bernoulli Scheme. Instead, we use an Autoencoder to learn an encoding that will partition the space while maintaining ergodicity. This means the learned trajectories on the discrete space will on average be consistent with those in dense continuous space, which reduces the memory and computation of what would otherwise involve solving high dimensional differential equations on a continuous domain. The learned symbolic encodings of the Autoencoder are closely correlated with the form of the latent world model itself.

3.1 Ornstein Isomorphism Theorem

The Ornstein isomorphism theorem is a deep result for ergodic theory. It states that if two different Bernoulli schemes have the same Kolmogorov entropy, then they are isomorphic. It reveals that many systems previously believed to be unrelated are in fact isomorphic; these include all finite stationary stochastic processes, including Markov chains and subshifts of finite type, Anosov flows and Sinai's billiards, ergodic automorphisms of the n-torus (uniform hyperbolic dynamics), and the continued fraction transform.

References

[1] Kahneman, Daniel. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.

[2] Bengio, Yoshua, From System 1 Deep Learning to System 2 Deep Learning, NeurIPS 2019.

[3] LeCun, Yann, A Tutorial on Energy-Based Learning, 2006.