Hamiltonian Generative Networks

1. Hamiltonian Mechanics: Introduction

All deterministic physical systems have a complete representation in terms of the position and momentum of each particle (a phase space).

We should try to express the dynamics of a mechanical system using the minimum representation possible that reflects the coordinate-invariance of the laws of physics. We should do this accounting for the fact that a mechanical system tries to optimize its 'action' from one moment to the next. Together this leads to the development of Hamiltonian Mechanics (reference: http://www.macs.hw.ac.uk/ simonm/mechanics.pdf).

2. Hamilton's Principle of Least Action: Extending Lagrangian Mechanics

Let the Lagrangian L be the difference of the kinetic and potential energies for a system, $L = T - V$ we define the action $A = A(q)$ from time t_1 to t_2 to be the functional

$$A(q) := \int_{t_1}^{t_2} L(q, \dot{q}, t)dt$$

The correct path of motion of a mechanical system with holonomic constraints and conservative external forces from time t_1 to t_2 is a stationary solution of the action. The correct path of motion $q = q(t)$ with $q = (q_1, ..., q_n)^T$ necessarily and sufficiently satisfies Lagrange's equations of motion for $j = 1, ..., n$:

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}_j}\Big) - \frac{\partial L}{\partial q_j} = 0$$

There are broad classes of problems that involve the minimization of quantities expressed under an integral. EX: Brachistochrome: Suppose a particle is allowed to slide freely along a wire under gravity from a point (x_1, y_1) to the origin $(0, 0)$. Find the curve $y = y(x)$ that minimizes the time of descent.

This is a special case of a general problem, the "variational problem". Suppose a function F is twice continuously differnetiable with respect to its arguments. Among all functions $y = y(x)$ which are twice continuously differentiable on the interval $[a, b]$ with $y(a)$ and $y(b)$ specified, find the one which extremizes the functional defined by

$$J(y) := \int_a^b F(x, y, y_x)dx$$

The function that extremizes the functional J satisfies the Euler-Lagrange equation on $[a, b]$ (since a differentiable functional is stationary at its local extrema),

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\Big(\frac{\partial F}{\partial u_x}\Big) = 0$$

We can use this to find the paths that minimize this action quantity and by this derive the dynamics of a mechanical system.

3. The Hamiltonian

The Hamiltonian Function is the Legendre transformation of the Lagrangian function,

$$H(q, p, t) := \dot{q} \cdot p - L(q, \dot{q}, t)$$

A hamiltonian function H induces a vector field on a phase space called the Hamiltonian vector field. This vector field describes ALL POSSIBLE DYNAMICS of the system. This corresponds to a Hamiltonian flow, a family of transformations of the phase space which (by Liouville's theorem) preserves the volume of the phase space. Suppose we have a Hamiltonian dynamical system with canonical coordinates q_i and conjugate momenta p_i where $i = 1, ..., n$. Then the phase space distribution $\rho(p, q)$ determines the probability $\rho(p, q)d^n q d^n p$ that the system will be found in the infinitesimal phase space volume $d^n q d^n p$. The Liouville equation governs the time evolution of $\rho(p, q; t)$,

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_{i=1}^{n} \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) = 0$$

This is the conservation of density in phase space. From Lagrange's equations above we can deduce the equations that govern the time evolution of our system, called Hamilton's Equations:

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q}, \frac{dq}{dt} = \frac{\partial H}{\partial p}$$

where $H(q, p, t)$ is the Hamiltonian. In a closed (conservative) system this is the sum of kinetic and potential energies T and V.

4. A General Approach to the Dynamics of Mechanical Systems

From the above we have a procedure to construct Hamilton's canonical equations.
1. Choose generalized coordinates $q = (q_1, ..., q_n)^T$ and construct $L(q, \dot{q}, t) = T - V$.
2. Define and compute the generalized momenta,

$$p_i = \frac{\partial L}{\partial \dot{q}_i}$$

and solve these relations to find $\dot{q}_i$.
3. Construct and compute the Hamiltonian function

$$H = \sum_{j=1}^{n} \dot{q}_j p_j - L$$

4. Write down equations of motion,

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \dot{p}_i = -\frac{\partial H}{\partial q_i}$$

5. 1D Hamiltonian of a particle of mass m

An example. The Hamiltonian represents the total energy of the system (the sum of kinetic and potential energy). Here our generalized coordinate q is just a position coordinate and p is the momentum mv, then

$$H = T + V, T = \frac{p^2}{2m}, V = V(q)$$

T time derivative of the momentum is Newtonian force $F = ma$ so from $\frac{dp}{dt} = -\frac{\partial H}{\partial q}$ tells us that force is the negative gradient of potential energy (as expected). The second Hamiltonian equation tells us that the particle's velocity is the derivative of kinetic energy wrt momentum.

6. Conservation Laws

Conservation laws pop out of this formalism as quantities that have time derivatives 0. Each of these are induced by some symmetry. This is a first step towards the more general Noether's theorems relating symmetries to conserved quantities.

7. Machine Learning and Hamiltonians

Any system capable of intelligent behavior within a dynamic environment requires a predictive model of the environment's dynamics.

As shown above, Hamiltonian mechanics describe dynamics in a way that is smooth, includes paths that respect conservation laws (morally, through symmetries), and have reversible time evolution.

Extend the image-manifold hypothesis by adding the Hamiltonian assumption: natural images lie on a low-dimensional manifold embedded within a high-dimensional pixel space and natural sequences of images trace out paths on this manifold that follow the equations of an abstract Hamiltonian. !!!

Can we learn a system's Hamiltonian from data? Can we infer a system's phase space from the observations available to machine learning systems?

8. Prior Work

ML approaches to dynamic modeling use discrete time steps and accumulate approximation errors when rolling out. This approach uses hamiltonian differential equations leading to slower divergence for longer rollouts. This approach also respects both invertibility and preserves the volume of the phase space which is both computationally preferable and more suitable for modeling real physics.

Hamiltonian Neural Network trains the gradients of a neural network to match the time derivative of a target system in a supervised fashion

$$L_{HNN} = \frac{1}{2}\Big[\Big(\frac{\partial H}{\partial p} - \frac{dq}{dt}\Big)^2 + \Big(\frac{\partial H}{\partial q} + \frac{dp}{dt}\Big)^2\Big]$$

but for this you need to know the true phase space! And the Hamiltonian is learned from the ground truth state space, not from pixel observations.

To learn with pixel observations you flatten the images then map using an encoder to a low dimensional $z = [q_t, p_t]$. And the dimension of this embedding is chosen from our ground truth knowledge of the phase space which is pepega. You also need a bunch of tricks. The latent embedding must be treated as an estimate of the position and momentum of the system and this is enforced as a constraint that encourages the time derivative of the position latent to equal the momentum latent using finite differences $L_{cc} = (p_t - (q_{t+1} - q_t))^2$ which is a contraint on the form of the momentum that only works for certain systems (like the pendulum they chose in their experiments...)

9. This Work: OV

Learning the Hamiltonian is in general extremely difficult and requires a lot of training. Here is presenting a general method for learning the hamiltonian from raw observations (such as pixels) by inferring a system's state with a generative model and then rolling it out with Hamilton's equations

10. Inference Network

Let the data $X = (x_0^1, ..., x_T^1), ..., (x_0^K, ..., x_T^K)$ be high dim noisy observations where each $x_i = G(s_i) = G(q_i)$ a non-deterministic function of the generalized position in the phase space. We want to infer the state and learn the Hamiltonian dynamics in phase space by observing sequences.

Inference takes a sequence of images $(x_0^i, ..., x_T^i)$ concatenated along the channel dimension and maps a posterior over the initial state $z \approx q_\phi(\cdot|x_0, ..., x_T)$ which corresponds to the system's coordinates in phase space at the first frame of the sequence. $q_\phi(z)$ is a diagonal gaussian with unit prior.

Then map samples from the posterior with another function $s_0 = f_\phi(z)$.

11. Hamiltonian Network

Parameterize a neural network with parameters γ that takes the inferred abstract state and maps it to a scalar $H_\gamma(s_t) \in \mathbb{R}$. Then we can use this to roll out the state space in time,

$$s_{t+1} = (q_{t+1}, p_{t+1}) = (q_t + \frac{\partial H}{\partial p_t} dt, p_t - \frac{\partial H}{\partial q_t} dt)$$

by Euler integration. But if the Hamiltonian is not separable there are more sophisticated ways to do this (which actually yield better results).

12. Decoder Network

Deconvolutional network that takes low dim representation and produces pixel reconstruction.

13. Objective Function

Given a sequence of $T + 1$ images, HGN optimizes,

$$L(\psi, \phi, \gamma, \theta; x_0, ..., x_T) = \frac{1}{T+1} \sum_{t=0}^{T} \left[\mathbb{E}_{q_\psi(z|x_1,..,x_t)} [\log p_{\phi,\gamma,\theta}(x_t|q_t)] \right] - KL(q_\psi(z)||p(z))$$

which is a time extended VAE objective with a reconstruction term for each frame and a KL term.

Key difference from standard VAE is the ability to produce rollouts through Hamilton's equations.

14. Comparison to Normalizing Flow

In a normalizing flow we have an invertible function implemented by a neural network f_i. In a Hamiltonian flow we learn the Hamiltonian by mapping $s_0 \to E$ by H_i then taking derivatives $\frac{\partial H}{\partial p}$ and $\frac{\partial H}{\partial q}$.

15. Learning Hamiltonian Flows

If instead of having one Hamiltonian shared across all time points we allow for a different Hamiltonian at each time point this gives us a flow.

Advantages:

1. They are invertible and volume preserving since they are Hamiltonians.
2. $s = (q, p)$ can be constrained. There are tricks in the Monte Carlo literature that allow you to treat momentum as a latent variable to get a tractable ELBO.