

# A Variational and Geometric Perspective on Learning Dynamics

Jamie Graham

2026

We present an integrated variational–geometric perspective on learning dynamics and identify structural principles shared across optimization, inference, and generative modeling.

## 1 Central Questions

- What structural principles govern the dynamics of learning, and how do variational and geometric ideas clarify their relationships?
- In what precise sense can learning dynamics be understood as variational flows of energy functionals under chosen geometric and stochastic structures?
- What is the minimal geometric and variational structure required to describe modern learning algorithms?
- Can deterministic optimization, stochastic gradient methods, variational inference, and diffusion be described within a common variational–geometric framework, and what structural elements differentiate them?

## 2 Core Sections

- Variational Mechanics
- Gradient Flows as Dissipative Variational Systems
- Stochastic Extensions and Thermodynamic Structure
- Distribution-Space Dynamics
- Connections to Optimization, VI, Diffusion

## 3 Variational Mechanics and Trajectory Optimization

### 3.1 Dynamical Systems

A dynamical system is a mathematical framework used to describe how something changes over time. It provides a rule that determines how the state of a system evolves, given its current

condition. The state represents the complete information needed to describe the system at a particular moment, such as the position of a particle, the configuration of a mechanical system, or the activity of a network. Once the system's initial state is specified, the dynamical system determines how that state develops into the future (and sometimes the past).

You can think of a dynamical system as consisting of two main ingredients: a space of possible states and a rule that describes how states change with time. As time progresses, the system follows a trajectory through this state space, representing its evolution.

In classical mechanics, it is helpful to distinguish **configuration** from **state**. A configuration is a point  $q \in Q$  recording “where the system is,” while the **mechanical state** is typically  $(q, \dot{q}) \in TQ$ . Velocities complete the state because mechanical laws are usually **second-order** in time: specifying  $q(t_0)$  alone does not determine the future, but specifying  $(q(t_0), \dot{q}(t_0))$  does (given the forces/constraints).

Formally, let  $M$  be a state space (often modeled as a smooth manifold). A continuous-time dynamical system is defined by a rule that assigns a trajectory to each initial state. This is typically specified by a differential equation of the form

$$\frac{dx}{dt} = F(x, t) \quad (1)$$

where  $x(t) \in M$  represents the state of the system at time  $t$ , and  $F$  is a function that determines how the state changes over time.

Equivalently, a dynamical system can be described by a flow

$$\phi_t : M \rightarrow M \quad (2)$$

such that  $\phi_t(x_0)$  gives the state of the system at time  $t$  starting from the initial state  $x_0$ , and satisfies

$$\phi_0(x) = x, \quad \phi_{t+s}(x) = \phi_t(\phi_s(x)). \quad (3)$$

Under standard regularity assumptions (e.g. the vector field  $F$  is locally Lipschitz in  $x$ ), these viewpoints are equivalent:  $F$  generates the flow  $\phi_t$  via solutions of the ODE, and  $\phi_t$  recovers  $F$  by differentiation at  $t = 0$ .

### 3.2 Configuration Space $Q$

In mechanics, a **Configuration Space**  $Q$  is a mathematical space where each point represents one complete way a system can be arranged or positioned. Instead of tracking objects directly in physical space, we describe the system using the smallest set of coordinates needed to uniquely specify its state.

Each set of coordinate values corresponds to a single point in configuration space, and as the system changes over time it traces a path through this space. The number of dimensions in configuration space equals the number of independent coordinates needed to describe the system. Thinking this way allows complex motion to be understood as movement through a geometric space of possibilities.

For example, a single particle moving in ordinary space can be fully described by three coordinates

$$(q_1, q_2, q_3) \quad (4)$$

so its configuration space is simply three-dimensional space  $\mathbb{R}^3$ , where each point represents one possible position of the particle.

### 3.3 Mechanical State Space $TQ$

For mechanical systems, the natural state space is not  $Q$  itself but instead a space called the **tangent bundle**  $TQ$ . Intuitively,  $TQ$  is the space of all pairs  $(q, v)$  where  $q \in Q$  and  $v \in T_q Q$  is a **tangent vector** at  $q$ , interpreted as a velocity.

This is why **Lagrangians live on  $TQ$** : at each time  $t$ , the velocity  $\dot{q}(t)$  is literally a tangent vector  $\dot{q}(t) \in T_{q(t)} Q$ , so  $L$  must accept both  $q$  and  $\dot{q}$  as inputs.

### 3.4 Dynamical Trajectories $q(t)$

A dynamical system evolves over time by moving through its configuration space. The function  $q(t)$  describes this evolution by specifying the system's configuration at each moment in time. In other words, for every time  $t$ , the value of  $q(t)$  tells you exactly where the system is within the space of all possible configurations.

As time varies,  $q(t)$  traces out a continuous curve through configuration space. This curve is called the system's **trajectory**, and it represents the full history of how the system moves from one configuration to another.

The quantity  $\dot{q}(t)$  describes how the system moves along this trajectory. It represents the **velocity through configuration space**, indicating both how quickly the configuration is changing and in what direction it is changing at time  $t$ . While  $q(t)$  specifies the system's position in configuration space,  $\dot{q}(t)$  describes how that position is evolving.

Together,  $q(t)$  and  $\dot{q}(t)$  describe both where the system is and how it is moving. We typically assume trajectories are at least continuously differentiable ( $C^1$ ) so that  $\dot{q}(t)$  exists.

Formally, let  $Q$  denote the configuration space of a system (typically modeled as a smooth manifold). A trajectory of the system is defined as a smooth time-parameterized curve

$$q : I \rightarrow Q \tag{5}$$

where  $I \subseteq \mathbb{R}$  is a time interval and  $q(t)$  specifies the system's configuration at time  $t$ .

The velocity of the system is defined as the time derivative of this curve. At each time  $t$ , the velocity is an element of the tangent space of  $Q$  at the point  $q(t)$ :

$$\dot{q}(t) = \frac{dq}{dt}(t) \in T_{q(t)} Q \tag{6}$$

where  $T_{q(t)} Q$  denotes the tangent space to the configuration space at  $q(t)$ .

Equivalently, the pair

$$(q(t), \dot{q}(t)) \tag{7}$$

defines a curve in the tangent bundle  $TQ$ , which is the space containing all configurations together with their associated velocities.

### 3.5 Functionals

A functional is a mathematical rule that takes an entire function as its input and produces a single number as its output. Instead of operating on individual numbers or vectors, a functional evaluates properties of whole curves, shapes, or fields. You can think of it as a way of assigning a numerical value to an entire path or configuration, often measuring quantities like total energy, length, or accumulated cost. Functionals are especially useful when studying systems where the object of interest is not a single point, but a continuous trajectory or distribution. In many areas of physics and mathematics, the behavior of a system can be determined by finding the function that makes a particular functional as small or large as possible.

For example, if  $y(x)$  is a curve between two points, the expression

$$J[y] = \int_a^b y(x)^2 dx \quad (8)$$

is a functional because it takes the entire function  $y(x)$  as input and outputs a single number representing the total accumulated squared value of the curve.

Formally, a functional is a mapping

$$J : \mathcal{F} \rightarrow \mathbb{R} \quad (9)$$

where  $\mathcal{F}$  is a space of functions (such as continuous or differentiable functions) and  $J$  assigns a real number to each function in that space.

### 3.6 Lagrangian Formalism

A Lagrangian  $L(q, \dot{q}, t)$  is a mathematical function that summarizes how a physical system moves by describing the balance between motion and stored energy. It is typically written as a function of a system's coordinates  $q$ , their rates of change  $\dot{q}$ , and time  $t$ , and it encodes the rules governing the system's dynamics. Instead of directly computing forces like in Newtonian mechanics, the Lagrangian allows us to determine motion by identifying the path a system takes between two configurations. This is done by constructing a quantity called the action and finding the path that makes this quantity **stationary** (often minimal in physical systems).

Lagrangians can be thought of as a **cost density** or a **scoring rule** that tells us how “favorable” or “natural” a system’s motion is at each moment in time. At its core often in classical mechanics, it measures a balance between two competing tendencies: kinetic energy  $T$ , which reflects how strongly a system is moving or spreading its motion, and potential energy  $V$ , which reflects how strongly the system is being pulled toward or held by its environment. The difference  $T - V$  can be thought of as measuring how freely the system is able to move relative to how constrained or trapped it is. If kinetic energy is large compared to potential energy, the system is moving freely; if potential energy dominates, the system is strongly restricted by forces or stored energy. You can imagine it like a traveler balancing momentum with terrain:  $T$  rewards motion and exploration, while  $V$  represents hills, valleys, or constraints that shape which routes are natural or costly. The system’s actual motion is determined by combining this balance over time into a total quantity called the action and selecting the path that makes this total score **stationary** (often minimal in physical systems). This viewpoint matters because it reframes “solving the equations of motion” as an optimization-style problem over trajectories. That is to say, variational mechanics reformulates dynamics as optimization over function spaces.

We will use the following **running example** throughout this section.

For a single particle of mass  $m$  moving in one dimension with position  $q(t)$  in a potential energy field  $V(q)$ , the Lagrangian is

$$L(q, \dot{q}, t) = T - V = \frac{1}{2}m\dot{q}^2 - V(q) \quad (10)$$

where  $\dot{q}$  is the particle's velocity,  $T$  is kinetic energy, and  $V$  is potential energy. Additionally, while classical mechanical systems often use  $T - V$ , modern variational formulations allow far more general Lagrangians.

Formally, a Lagrangian is a function

$$L : TQ \times \mathbb{R} \rightarrow \mathbb{R} \quad (11)$$

where  $Q$  is the configuration space of the system,  $TQ$  is its tangent bundle (the space of coordinates  $q$  and velocities  $\dot{q}$ ), and  $L(q, \dot{q}, t)$  assigns a real number to each configuration, velocity, and time.

### 3.7 Action Functional

We now introduce the central object linking **dynamics** and **optimization**. The **action functional** assigns a single number to an entire trajectory. Instead of evaluating the Lagrangian at one instant, we integrate  $L(q, \dot{q}, t)$  over time between two fixed moments  $t_0$  and  $t_1$ . That integral measures the total “cost” or “score” accumulated along the path. Physical trajectories are those that make this total score **stationary**—small variations of the path do not change the value of the action to first order. So the action functional turns the question “what path does the system follow?” into an optimization problem over curves: nature (or an algorithm) chooses the path that makes  $S$  stationary over all curves in a space of infinite dimensional functions connecting the same endpoints.

$$S[q(\cdot)] = \int_{t_0}^{t_1} L(q(t), \dot{q}(t), t) dt \quad (12)$$

Where:

- $q(t)$  is a trajectory in configuration space  $Q$
- $\dot{q}(t)$  is velocity (a tangent vector in  $T_{q(t)}Q$ )
- $L$  is the Lagrangian

Typically, one varies over trajectories satisfying boundary conditions  $q(t_0) = q_0$  and  $q(t_1) = q_1$ , and seeks paths that make  $S$  **stationary** (often minimal in physical systems). In our running example, substituting  $L(q, \dot{q}, t) = \frac{1}{2}m\dot{q}^2 - V(q)$  makes  $S[q(\cdot)]$  an explicit “score” assigned to entire paths.

### 3.8 Toward the Euler–Lagrange Equation

Knowing that physical trajectories are those that make the action stationary is not yet enough to *compute* them. We need a **local** condition—a differential equation—that a curve must satisfy at each time if it is to extremize  $S$ . That condition is the **Euler–Lagrange equation**.

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \frac{\partial L}{\partial \dot{x}} = 0 \quad (13)$$

We will motivate that this condition emerges naturally from the stationarity condition on the action functional. Consider this as an analogue of setting the derivative to zero in ordinary calculus: instead of “derivative of a function equals zero at a critical point,” we get “a certain combination of derivatives of  $L$  evaluated along the curve equals zero at each time.”

Our derivation proceeds by **variation**. We take a candidate trajectory  $q(t)$  and perform a first-order perturbation of it slightly to  $q(t) + \epsilon\eta(t)$ , where  $\eta(t)$  is an arbitrary smooth function that vanishes at the endpoints, giving the boundary condition:

$$\eta(t_0) = \eta(t_1) = 0 \quad (14)$$

Thus we only consider curves that still connect the same two configurations  $q_0$  and  $q_1$ .

We then ask: for  $S[q(t)]$  to be stationary at  $q(t)$ , the first-order change in  $S$  with respect to  $\epsilon$  must vanish for *every* such perturbation  $\eta$ , that is:

$$\delta S = \frac{d}{d\epsilon} S[q(t) + \epsilon\eta(t)] \Big|_{\epsilon=0} = 0 \quad (15)$$

Let us also insert our first order perturbation of the trajectory  $q(t)$  into the action functional.

$$S[q(t) + \epsilon\eta(t)] = \int_{t_0}^{t_1} L(q(t) + \epsilon\eta(t), \dot{q}(t) + \epsilon\dot{\eta}(t), t) dt \quad (16)$$

Altogether, this gives us

$$\delta S = \int \left( \frac{\partial L}{\partial q} \eta + \frac{\partial L}{\partial \dot{q}} \dot{\eta} \right) dt = 0 \quad (17)$$

This shows that the variation of the action functional is separable into a term that encodes sensitivity to position as well as sensitivity to velocity, hinting that motion depends on both.

We now perform a trick on the second term, using integration by parts.

$$\int \left( \frac{\partial L}{\partial \dot{q}} \dot{\eta} \right) dt \quad (18)$$

Integrating by parts:

$$= \left[ \frac{\partial L}{\partial \dot{q}} \eta(t) \right]_{t_0}^{t_1} - \int \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}} \right) \eta(t) dt \quad (19)$$

But notice that the first term vanishes due to the aforementioned boundary conditions.  $\eta(t_0) = \eta(t_1) = 0$

So we obtain:

$$\delta S = \int \left( \frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right) \eta(t) dt = 0 \quad (20)$$

Note now that the variation of  $S$  collapses into a single term multiplied by an arbitrary smooth function  $\eta(t)$ , yet in order for the action to be stationary for all admissible variations, the integral must be equal to zero. Since  $\eta(t)$  is arbitrary, the only way this is possible is if the condition

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} = 0 \quad (21)$$

which is the Euler-Lagrange equation, so we are done.

$$\square \quad (22)$$

Intuitively, we see that the Euler-Lagrange equation emerges as a local constraint on the Lagrangian, following from extremizing the action functional  $S$ .

### 3.9 On The Implications of Stationary Action

What does stationary action really imply?

$$\delta S = 0 \quad (23)$$

This means that the physical trajectory  $q(t)$  that emerges in a system is a stationary point in function space. This is not necessarily a minimum, it could be a maximum or saddle point, but it is a stationary point.

### 3.10 An Example: Recovering Newton's Second Law from Stationary Action

A single particle of mass  $m$  in one dimension with position  $x(t)$  in a potential  $V(x)$  has Lagrangian

$$L = \frac{1}{2}m\dot{x}^2 - V(x). \quad (24)$$

Applying the Euler-Lagrange equation

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \frac{\partial L}{\partial \dot{x}} = 0 \quad (25)$$

In one dimension, we note that the partial derivative of the Lagrangian with respect to  $x$  is actually equal to the negative gradient of the potential.

$$\frac{\partial L}{\partial x} = -\frac{\partial V}{\partial x} = -\nabla V \quad (26)$$

Note that taking the derivative of the lagrangian with respect to velocity gives

$$\frac{\partial L}{\partial \dot{x}} = m\dot{x} \quad (27)$$

Taking a time derivative gives

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{x}} = m\ddot{x} \quad (28)$$

Substituting Euler-Lagrange, we get

$$m\ddot{x} = -\nabla V = F \quad (29)$$

Which is Newton's second law. Note that Newton's second law thus emerges from the Euler-Lagrange equation, which emerges from stationary action.

Local dynamics (the ODE at each instant) emerges from global optimization (extremizing  $S$  over entire trajectories). This is a central insight we can apply to modern settings, where trajectory selection by a variational principle appears in learning and diffusion.

## 4 Gradient Flows as Dissipative Variational Systems

### 4.1 From Stationary Action to Energy Dissipation

In classical variational mechanics, trajectories arise from the stationarity of an action functional. Under suitable regularity assumptions, this yields Euler–Lagrange equations whose associated dynamics are typically conservative. In particular, for time-independent Lagrangians, the Hamiltonian energy

$$H(q, \dot{q}) \tag{30}$$

is preserved along trajectories:

$$\frac{d}{dt} H(q(t), \dot{q}(t)) = 0. \tag{31}$$

The variational principle therefore selects trajectories consistent with energy conservation rather than energy decay.

Learning dynamics, by contrast, are intrinsically dissipative. Let

$$\mathcal{E} : \Theta \rightarrow \mathbb{R} \tag{32}$$

denote a loss or objective functional defined on a parameter space  $\Theta$ . A fundamental property of most training procedures is that along the trajectory  $\theta(t)$ , the objective decreases:

$$\frac{d}{dt} \mathcal{E}(\theta(t)) \leq 0. \tag{33}$$

Rather than preserving an energy functional, learning algorithms are designed to drive the system toward lower-energy configurations.

This structural distinction suggests that stationary action is not the appropriate variational framework for learning. Instead, learning dynamics are more naturally described as gradient flows of energy functionals, in which dissipation is fundamental rather than accidental.

In the context of supervised learning, the energy functional  $\mathcal{E}$  may represent empirical risk,

$$\mathcal{E}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(f_\theta(x_i), y_i), \tag{34}$$

where  $f_\theta$  is a parameterized model and  $\ell$  a loss function. More generally,  $\mathcal{E}$  may denote any differentiable objective functional whose minimization defines the learning problem.

### 4.2 Gradient Flow in Euclidean Parameter Space

A gradient flow is a continuous-time dynamical system that evolves according to the direction of steepest descent of an energy functional.

The gradient flow of an energy functional is defined by the following differential equation:

$$\dot{\theta}(t) = -\nabla \mathcal{E}(\theta(t)). \tag{35}$$

Taking the time derivative of the energy functional, we get:

$$\frac{d}{dt}\mathcal{E}(\theta(t)) = \langle \nabla\mathcal{E}(\theta(t)), \dot{\theta}(t) \rangle \quad (36)$$

Substituting the gradient flow equation into the time derivative of the energy functional, we get:

$$\frac{d}{dt}\mathcal{E}(\theta(t)) = -\langle \nabla\mathcal{E}(\theta(t)), \nabla\mathcal{E}(\theta(t)) \rangle = -|\nabla\mathcal{E}(\theta(t))|^2 \leq 0 \quad (37)$$

Thus the energy decreases monotonically along trajectories of the flow. In contrast to conservative Euler–Lagrange dynamics, where stationarity of an action yields energy preservation, gradient flow dynamics are intrinsically dissipative: the energy functional itself drives motion toward lower-energy configurations.

Importantly, the definition of the gradient depends on the underlying metric structure on  $\Theta$ . The Euclidean inner product induces the standard gradient operator; alternative choices of geometry give rise to distinct gradient flows.

### 4.3 Minimizing Movements: A Dissipative Variational Principle

Although gradient flow is defined as a differential equation, it admits an equivalent variational characterization. This formulation restores variational structure in a manner fundamentally distinct from stationary action.

Let  $\eta > 0$  be a time-step parameter. Given a current state  $\theta_k \in \Theta$ , define the next state by the minimization problem

$$\theta_{k+1} = \arg \min_{\theta \in \Theta} \left\{ \mathcal{E}(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|^2 \right\}. \quad (38)$$

This scheme selects the subsequent state by balancing two competing effects:

- reduction of the energy  $\mathcal{E}$ ,
- proximity to the previous state  $\theta_k$ .

Unlike stationary action, which selects entire trajectories through a global extremization principle, the minimizing movement scheme determines evolution incrementally through a sequence of local variational problems over states.

As  $\eta \rightarrow 0$ , the piecewise-linear interpolation of the discrete iterates converges, under suitable regularity assumptions, to the solution of the gradient flow equation

$$\dot{\theta}(t) = -\nabla\mathcal{E}(\theta(t)). \quad (39)$$

In this sense, gradient flow may be understood as the continuous-time limit of a time-discrete variational principle. Dissipation therefore arises not from stationary action over trajectories, but from successive minimization of energy penalized by a metric cost of motion.

Stationary action extremizes a functional over curves in configuration space, whereas minimizing movements extremize a functional over successive states in metric space. The former produces conservative dynamics; the latter generates dissipative evolution.

## 4.4 Gradient Flow on a Riemannian Manifold

The notion of gradient depends fundamentally on the underlying metric structure of the state space. To make this dependence explicit, let  $(M, g)$  be a Riemannian manifold and let  $\mathcal{E} : M \rightarrow \mathbb{R}$  be a smooth energy functional.

At each point  $\theta \in M$ , the differential  $d\mathcal{E}_\theta$  defines a covector in the cotangent space  $T_\theta^*M$ . The Riemannian metric  $g$  induces an identification between tangent and cotangent spaces, allowing one to define the Riemannian gradient  $\nabla_g \mathcal{E}$  implicitly by

$$g_\theta(\nabla_g \mathcal{E}, v) = d\mathcal{E}_\theta(v) \quad \text{for all } v \in T_\theta M. \quad (40)$$

The gradient flow of  $\mathcal{E}$  with respect to the metric  $g$  is then defined by

$$\dot{\theta}(t) = -\nabla_g \mathcal{E}(\theta(t)). \quad (41)$$

Along any sufficiently smooth solution of this equation, the energy evolves according to

$$\frac{d}{dt} \mathcal{E}(\theta(t)) = d\mathcal{E}_\theta(\dot{\theta}) = -g_\theta(\nabla_g \mathcal{E}, \nabla_g \mathcal{E}) = -\|\nabla_g \mathcal{E}\|_g^2 \leq 0. \quad (42)$$

Thus the dissipative character of the dynamics is preserved under arbitrary choices of Riemannian geometry: the energy decreases at a rate determined by the squared norm of its gradient with respect to the metric  $g$ .

## 4.5 Conservative and Dissipative Variational Structures

The preceding sections exhibit two distinct variational paradigms. Classical mechanics arises from stationarity of an action functional over trajectories and yields conservative dynamics. Learning dynamics, by contrast, arise from gradient flow of an energy functional and yield dissipative evolution.

The structural differences may be summarized as follows:

Conservative Dynamics	Dissipative Dynamics
Action functional over trajectories	Energy functional over states
Euler–Lagrange / Hamiltonian equations	Gradient flow equations
Second-order in time	First-order in time
Energy preserved ( $\frac{dH}{dt} = 0$ )	Energy decreases ( $\frac{d\mathcal{E}}{dt} \leq 0$ )
Symplectic structure	Metric structure

Conservative systems preserve geometric structure and exhibit reversible evolution. Dissipative systems contract energy and generically exhibit irreversible behavior. While both arise from variational principles, the nature of the extremization differs fundamentally: stationary action selects entire trajectories, whereas gradient flow emerges from successive minimization of energy relative to a metric cost of motion.