Variational Flows in Learning: Geometry, Thermodynamics, and Distribution-Space Dynamics

Jamie Graham

jgrah52@uwo.ca
University of Western Ontario

March 2026

Abstract

We present a variational–geometric perspective on learning dynamics that highlights a small set of structural ingredients shared across optimization, inference, and generative modeling. We begin by contrasting conservative dynamics arising from stationary action in classical mechanics with the intrinsically dissipative dynamics of learning, modeled as gradient flows of an energy functional on parameter space. We then introduce stochastic extensions via Langevin-type dynamics, which lift parameter trajectories to evolving laws and endow learning with thermodynamic structure: an invariant Gibbs measure, a Fokker–Planck description in distribution space, and a free-energy Lyapunov functional exhibiting monotone dissipation. Building on this, we formulate distribution-space dynamics on $\mathcal{P}(\Theta)$ and emphasize that "steepest descent" depends on the chosen geometry, focusing on Wasserstein (transport) and KL/Fisher (information) structures and their associated minimizing-movement schemes. Finally, we show how deterministic optimization, stochastic gradient methods, variational inference, and diffusion models arise as specializations corresponding to different choices of state variable (parameters vs. measures), functional (energy, free energy, KL), geometry (Euclidean/Riemannian, Wasserstein, KL/Fisher), and stochasticity. The resulting framework does not identify learning with mechanics, but organizes modern learning algorithms as variational flows under prescribed geometric and thermodynamic structure, suggesting a principled lens for analyzing and designing optimizers and generative procedures. This survey is intended to be self-contained and does not assume prior exposure to optimal transport or information geometry.

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?
- 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 Outline

We first review variational mechanics and the Euler–Lagrange framework (Section 3), then contrast it with gradient flows as dissipative variational systems (Section 4). Stochastic extensions via Langevin dynamics and their thermodynamic structure are developed in Section 5, followed by distribution-space dynamics on $\mathcal{P}(\Theta)$ and the role of Wasserstein versus KL geometry (Section 6). Section 7 connects these structures to deterministic optimization, stochastic gradient methods, variational inference, and diffusion models.

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

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

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

$$\phi_t : M \to M \tag{2}$$

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

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

Under standard regularity assumptions (e.g. the vector field F is locally Lipschitz in x), these viewpoints are equivalent: F generates the flow ϕ_t via solutions of the ODE, and ϕ_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

$$(q_1, q_2, q_3) \tag{4}$$

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 $v \in T_qQ$ 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

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

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

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

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

$$J : \mathcal{F} \to \mathbb{R} \tag{9}$$

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 [3, 5].

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

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

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

$$L : TQ \times \mathbb{R} \to \mathbb{R} \tag{11}$$

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.

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

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**:

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = 0. \tag{13}$$

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:

$$\eta(t_0) = \eta(t_1) = 0 \tag{14}$$

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 ϵ must vanish for *every* such perturbation η, that is:

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

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

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

Altogether, this gives us

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

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.

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

Integrating by parts:

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

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

So we obtain:

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

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

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} = 0 \tag{21}$$

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

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?

$$\delta S = 0 \tag{22}$$

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

$$L = \frac{1}{2} m \dot{x}^2 - V(x). \tag{23}$$

Applying the Euler–Lagrange equation

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \frac{\partial L}{\partial \dot{x}} = 0 \tag{24}$$

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.

$$\frac{\partial L}{\partial x} = -\frac{\partial V}{\partial x} = -\nabla V \tag{25}$$

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

$$\frac{\partial L}{\partial \dot{x}} = m\dot{x} \tag{26}$$

Taking a time derivative gives

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = m\ddot{x} \tag{27}$$

Substituting Euler-Lagrange, we get

$$m\ddot{x} = -\nabla V = F \tag{28}$$

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 the previous section, the system's path was chosen by making a single number—the action—stationary, and energy was conserved. In learning, we want the opposite: we want a quantity (the loss) to *decrease* over time.

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. The **Hamiltonian** $H(q, \dot{q})$ is the total energy; for our running example it is $T + V$. In particular, for time-independent Lagrangians, H, defined by $H = p \cdot \dot{q} - L$ with momentum $p = \partial L/\partial \dot{q}$ (or equivalently $H = T + V$ for standard mechanical Lagrangians), is preserved along trajectories:

$$\frac{d}{dt} H(q(t), \dot{q}(t)) = 0. \tag{29}$$

The variational principle therefore selects trajectories consistent with energy conservation rather than energy decay. Nature picks a path that conserves energy; training picks a path that drains the loss.

Learning dynamics, by contrast, are intrinsically dissipative. Let

$$\mathcal{E} : \Theta \to \mathbb{R} \tag{30}$$

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

$$\frac{d}{dt}\mathcal{E}(\theta(t)) \leq 0. \tag{31}$$

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,

$$\mathcal{E}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i), \tag{32}$$

where f_θ is a parameterized model and ℓ a loss function. So $\mathcal{E}(\theta)$ is the average error on the training set; we want to drive it down. 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 the continuous-time version of "always take a small step in the direction that decreases the energy the most." It is a continuous-time dynamical system that evolves according to the direction of steepest descent of an energy functional.

In calculus, steepest descent is the direction opposite to the gradient; here we turn that into an ODE. The gradient flow of an energy functional is defined by the following differential equation:

$$\dot{\theta}(t) = -\nabla\mathcal{E}(\theta(t)). \tag{33}$$

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

$$\frac{d}{dt}\mathcal{E}(\theta(t)) = \langle \nabla\mathcal{E}(\theta(t)), \dot{\theta}(t)\rangle \tag{34}$$

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

$$\frac{d}{dt}\mathcal{E}(\theta(t)) = -\langle \nabla\mathcal{E}(\theta(t)), \nabla\mathcal{E}(\theta(t))\rangle = -|\nabla\mathcal{E}(\theta(t))|^2 \leq 0 \tag{35}$$

So the energy never increases—it can only stay constant or decrease. 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 Θ. The Euclidean inner product induces the standard gradient operator; alternative choices of geometry give rise to distinct gradient flows. For now we use the usual Euclidean notion of length and angle.

4.3 Minimizing Movements: A Dissipative Variational Principle

We can also describe gradient flow without writing an ODE: at each time step, choose the next point by making a single *minimization* that balances "get lower energy" and "don't move too far." Although gradient flow is defined as a differential equation, it admits an equivalent variational characterization [2]. This formulation restores variational structure in a manner fundamentally distinct from stationary action.

Think of it as: from θ_k, where should I step so that I reduce $\mathcal{E}$ without jumping too far in one go? Let $\eta > 0$ be a time-step parameter. Given a current state $\theta_k \in \Theta$, define the next state by the minimization problem

$$\theta_{k+1} = \arg\min_{\theta \in \Theta} \left\{ \mathcal{E}(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|^2 \right\}. \tag{36}$$

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

- reduction of the energy $\mathcal{E}$,
- proximity to the previous state θ_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 $\eta \to 0$, the piecewise-linear interpolation of the discrete iterates converges, under suitable regularity assumptions, to the solution of the gradient flow equation

$$\dot{\theta}(t) = -\nabla \mathcal{E}(\theta(t)). \tag{37}$$

So the discrete "minimize energy plus penalty" steps blend into the continuous gradient flow.

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 picks a whole curve; minimizing movements pick the next point one step at a time. 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

So far "steepest" meant steepest in the usual Euclidean sense. If we change how we measure lengths and angles (a Riemannian metric), the direction of steepest descent changes. 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 $\mathcal{E} : M \to \mathbb{R}$ be a smooth energy functional.

The derivative of $\mathcal{E}$ gives a covector (a linear map on tangent vectors); the metric turns that covector into a tangent vector, which we call the gradient. 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

$$g_\theta(\nabla_g \mathcal{E}, v) = d\mathcal{E}_\theta(v) \quad \text{for all } v \in T_\theta M. \tag{38}$$

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

$$\dot{\theta}(t) = -\nabla_g \mathcal{E}(\theta(t)). \tag{39}$$

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

$$\frac{d}{dt}\mathcal{E}(\theta(t)) = d\mathcal{E}_\theta(\dot{\theta}) = -g_\theta(\nabla_g \mathcal{E}, \nabla_g \mathcal{E}) = -\|\nabla_g \mathcal{E}\|_g^2 \leq 0. \tag{40}$$

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. (On the sphere, for example, the Riemannian gradient is the projection of the Euclidean gradient onto the tangent plane.)

4.5 Conservative and Dissipative Variational Structures

We now summarize the two variational pictures side by side: one for mechanics (conservative) and one for learning (dissipative). 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

So the same idea—extremize a functional—leads to opposite behavior depending on *what* we extremize and *over what* (curves vs single states).

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.

5 Stochastic Extensions and Thermodynamic Structure

So far we have imagined a single trajectory sliding downhill. Real training is noisy; this section makes that noise explicit and shows it gives the system a thermodynamic structure: an equilibrium

distribution (Gibbs measure), a PDE describing how probability mass evolves (Fokker–Planck), and a quantity that always decreases (free energy). These ideas underpin both stochastic optimization and, later, variational inference and diffusion models.

5.1 Why Stochasticity? Fluctuations, Finite Data, and Exploration

Deterministic gradient flows provide an idealized continuous-time description of learning as dissipative motion in parameter space. However, modern training procedures are intrinsically stochastic for at least three reasons.

- **Minibatch gradients.** In large-scale learning, the loss $\mathcal{E}(\theta)$ is often a sum over many data points. Computing the full gradient is costly, so we estimate it from a random subset (a minibatch). That estimate is noisy—it fluctuates around the true gradient—so the effective dynamics are stochastic.
- **Finite data.** Even with the full dataset, the empirical risk is only an approximation to the true population risk. Sampling uncertainty and model misspecification introduce randomness that is often modeled as effective noise on the parameters.
- **Exploration.** In nonconvex settings, purely deterministic descent can get stuck in local minima or saddle regions. Injecting noise allows the system to "kick" out of shallow basins and explore the energy landscape, often improving generalization and convergence to flatter minima.

A natural way to add noise is to replace the ODE by a stochastic differential equation: the same drift term plus a random kick at each instant. A natural mathematical extension therefore replaces the ODE

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

by a **stochastic differential equation (SDE)** on parameter space:

$$d\theta_t = b(\theta_t, t)\, dt + \sigma(\theta_t, t)\, dW_t, \tag{42}$$

where b is a **drift** (deterministic push), σ controls the strength of random kicks, and W_t is a **Wiener process** (standard Brownian motion). Informally, dW_t is a tiny random nudge whose size is of order $\sqrt{dt}$; over time these increments add up to continuous but nowhere differentiable paths. Rigorous treatment requires stochastic calculus (Itô or Stratonovich); for our purposes it suffices to think of the SDE as "gradient descent plus random noise whose magnitude is set by σ."

This formulation introduces a **dual viewpoint**: one may study (i) *individual sample paths* $\theta_t(\omega)$ (a single run of the process) and (ii) the *evolution of the law* ρ_t of θ_t on Θ (how the distribution over parameters changes over time). The second viewpoint will lead to a PDE for the density ρ_t and to gradient flows on the space of measures.

5.2 Overdamped Langevin Dynamics as Noisy Gradient Flow

The standard stochastic version of gradient descent is **overdamped Langevin dynamics**: gradient drift plus white noise [9]:

$$d\theta_t = -\nabla \mathcal{E}(\theta_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t, \tag{43}$$

where β^{-1} acts like a temperature: larger β^{-1} means more noise. When $\beta \to \infty$, the noise vanishes and (43) formally reduces to deterministic gradient flow; when β is finite, the dynamics balance descent with random kicks.

Interpretation. The drift term $-\nabla\mathcal{E}(\theta_t)\,dt$ drives the system toward lower energy (dissipation), while $\sqrt{2\beta^{-1}}\,dW_t$ injects unbiased fluctuations (exploration). The factor $\sqrt{2\beta^{-1}}$ is chosen so that, as we will see, the equilibrium distribution has a clean Gibbs form. The resulting dynamics are *irreversible*: they do not preserve phase-space volume and naturally admit a thermodynamic structure with a distinguished equilibrium state.

Example (one dimension). Let $\Theta = \mathbb{R}$ and $\mathcal{E}(\theta) = \frac{1}{2}\theta^2$ (a quadratic well). Then (43) reads $d\theta_t = -\theta_t\,dt + \sqrt{2\beta^{-1}}\,dW_t$. The drift pulls θ toward zero; the noise spreads it. So the drift pulls toward zero while the noise spreads the distribution; at long times they balance at the Gibbs distribution—the distribution of θ_t converges to a Gaussian centered at zero with variance β^{-1}, which is exactly the Gibbs measure $Z^{-1}e^{-\beta\mathcal{E}(\theta)}$ for this $\mathcal{E}$.

5.3 Stationary Measures and the Gibbs Distribution

If we run Langevin forever, the distribution of θ_t typically settles down to a fixed distribution ρ_∞ that no longer changes in time. Under suitable regularity and growth assumptions on $\mathcal{E}$ (e.g., $\mathcal{E}$ smooth and growing sufficiently at infinity so that $Z < \infty$), Langevin dynamics admits a unique **invariant** (stationary) measure: if $\theta_0 \sim \rho_\infty$, then $\theta_t \sim \rho_\infty$ for all $t \geq 0$. That equilibrium distribution has a standard form from statistical physics: it is proportional to $e^{-\beta\mathcal{E}(\theta)}$. That measure has the **Gibbs** (or **Boltzmann**) form:

$$\rho_\infty(\theta) = Z^{-1}e^{-\beta\mathcal{E}(\theta)}, \qquad Z = \int_\Theta e^{-\beta\mathcal{E}(\theta)}\,d\theta. \tag{44}$$

The normalizing constant Z is the **partition function**; it ensures $\int \rho_\infty(\theta)\,d\theta = 1$. So Z is just the constant that makes the total probability equal to 1.

Intuition. The stochastic flow does not (in general) converge to a single minimizer; rather it approaches a *distribution* concentrated near low-energy regions. The parameter β (inverse temperature) controls the tradeoff: large β (low temperature) sharpens ρ_∞ around minima, so the process spends most of its time near the lowest energy; small β (high temperature) spreads the distribution, so higher-energy regions retain more probability mass. In the limit of zero temperature ($\beta \to \infty$), all mass concentrates at the minimizer(s) of $\mathcal{E}$; in that sense "cooling" recovers deterministic optimization. This equilibrium picture is central to both statistical physics and Bayesian inference: the Gibbs measure is the posterior in many statistical models when $\mathcal{E}$ is the negative log-posterior.

5.4 From Sample Paths to Densities: The Fokker–Planck Equation

So far we followed a single random trajectory θ_t. Equally important is how the *probability density* of θ_t evolves—that is described by a deterministic PDE. Its **law** is the probability density $\rho_t(\theta)$ such that $\mathbb{P}(\theta_t \in A) = \int_A \rho_t(\theta)\,d\theta$. How does ρ_t change in time?

The density ρ_t associated with the Langevin SDE (43) satisfies the **Fokker–Planck** (or **Kolmogorov forward**) equation [13]:

$$\partial_t\rho_t = \nabla\cdot(\rho_t\nabla\mathcal{E}) + \beta^{-1}\Delta\rho_t. \tag{45}$$

So the Fokker–Planck equation is the continuity equation for the density of a cloud of particles following the SDE. Derivation (sketch): the SDE describes how individual particles move; the flux of probability has two contributions: a **drift** term (mass is advected by the vector field $-\nabla\mathcal{E}$) and a **diffusion** term (mass spreads by Brownian motion). Collecting the divergence of the total flux yields (45).

This PDE makes explicit the two competing effects:

- **Drift** toward low energy: $\nabla\cdot(\rho_t\nabla\mathcal{E})$ describes probability mass being pushed in the direction $-\nabla\mathcal{E}$, i.e., downhill on the energy landscape. This term alone would concentrate ρ_t at minima.
- **Diffusion**: $\beta^{-1}\Delta\rho_t = \beta^{-1}\nabla\cdot(\nabla\rho_t)$ is the Laplacian of the density; it tends to flatten ρ_t and increase entropy. It opposes the concentration induced by the drift.

Drift concentrates mass downhill; diffusion spreads it. At equilibrium they balance, and $\partial_t\rho_t = 0$ yields $\rho_\infty \propto e^{-\beta\mathcal{E}}$ as in (44).

5.5 Free Energy as a Lyapunov Functional

There is a single quantity on the space of densities that always decreases along the Fokker–Planck flow: the free energy. It plays the role of a Lyapunov function. A central thermodynamic quantity is the **(Helmholtz) free energy** functional on densities. Free energy has two parts: expected energy (penalizes mass on high-$\mathcal{E}$ regions) and an entropy term (penalizes overly concentrated densities):

$$\mathcal{F}(\rho) = \int_\Theta \mathcal{E}(\theta)\,\rho(\theta)\,d\theta + \beta^{-1}\int_\Theta \rho(\theta)\log\rho(\theta)\,d\theta. \tag{46}$$

- The first term is the **expected energy** $\mathbb{E}_\rho[\mathcal{E}]$: it penalizes placing probability mass on high-energy regions.
- The second term is β^{-1} times the **negative entropy** $-\int\rho\log\rho$ (or entropy with a sign convention). Entropy is maximized by a uniform distribution; so this term penalizes overly concentrated ρ and encourages spreading.

The competition between "low energy" and "high entropy" formalizes the exploration–exploitation tradeoff: the Gibbs measure ρ_∞ is precisely the minimizer of $\mathcal{F}$, balancing concentration near minima with dispersion. So the Gibbs measure is the unique equilibrium both for the SDE and for the free energy.

Equivalently, $\mathcal{F}$ can be written in terms of the Gibbs measure ρ_∞:

$$\mathcal{F}(\rho) - \mathcal{F}(\rho_\infty) = \beta^{-1}\,\mathrm{KL}(\rho\,\|\,\rho_\infty), \tag{47}$$

where $\mathrm{KL}(\rho\|\rho_\infty) = \int\rho\log(\rho/\rho_\infty)\,d\theta$ is the **Kullback–Leibler divergence** [8]. So minimizing free energy is the same as driving ρ toward ρ_∞.

Energy dissipation in distribution space. Along solutions of the Fokker–Planck equation (45), the free energy decreases monotonically:

$$\frac{d}{dt}\mathcal{F}(\rho_t) \le 0, \tag{48}$$

with a precise dissipation identity (under suitable regularity) of the form

$$\frac{d}{dt}\mathcal{F}(\rho_t) = -\int_\Theta \rho_t(\theta)\, \left\|\nabla\left(\mathcal{E}(\theta) + \beta^{-1}\log\rho_t(\theta)\right)\right\|^2 d\theta \le 0. \tag{49}$$

The integrand is the squared norm of the **score** (or chemical potential gradient); it is zero only when $\rho_t = \rho_\infty$. So the system relaxes to ρ_∞ by flowing downhill in free energy. Thus Langevin dynamics may be viewed as a *gradient flow of free energy* in an appropriate geometry on probability measures—a perspective we make explicit in the next section.

5.6 A Variational Viewpoint: Minimizing Movements for Measures

Just as gradient flow in parameter space had a minimizing-movement formulation, the Fokker–Planck flow can be obtained by a sequence of minimizations over probability measures. The minimizing movement scheme of Section 4 extends naturally to the space of probability measures. In particular, the Fokker–Planck equation (45) arises as the continuous-time limit of the **Jordan–Kinderlehrer–Otto (JKO)** scheme [7]. Here W_2 measures the cost of rearranging one distribution into another by moving mass; it is the "earth mover's" distance:

$$\rho_{k+1} = \arg\min_\rho \left\{\mathcal{F}(\rho) + \frac{1}{2\eta}W_2^2(\rho,\rho_k)\right\}, \tag{50}$$

At each step we choose the next distribution ρ_{k+1} by minimizing free energy $\mathcal{F}(\rho)$ while penalizing how "far" ρ is from ρ_k in transport cost. As the time step tends to zero, these discrete updates converge to the Fokker–Planck equation, so diffusion is steepest descent of free energy in Wasserstein geometry. This provides a purely variational characterization of thermodynamic relaxation, directly paralleling minimizing movements in parameter space.

5.7 Connection to Learning: SGD as Approximate Langevin Dynamics

Stochastic gradient descent uses a noisy gradient; under certain conditions, that noise can be approximated by a Langevin-type SDE. SGD updates

$$\theta_{k+1} = \theta_k - \eta\,\widehat{\nabla\mathcal{E}}(\theta_k) \tag{51}$$

use a noisy gradient estimator $\widehat{\nabla\mathcal{E}}$ (e.g., the gradient of the loss on a random minibatch). Write $\widehat{\nabla\mathcal{E}}(\theta) = \nabla\mathcal{E}(\theta) + \xi$, where ξ is the error. In regimes where the minibatch is large enough that ξ is approximately Gaussian with covariance $\Sigma(\theta)$, the continuous-time limit of SGD (with step size $\eta \to 0$ and appropriate scaling) is often approximated by an SDE of Langevin type [16]:

$$d\theta_t \approx -\nabla\mathcal{E}(\theta_t)\,dt + (\eta\Sigma(\theta_t))^{1/2}\,dW_t. \tag{52}$$

So SGD behaves like Langevin dynamics with a *parameter-dependent* noise strength. So SGD both minimizes the loss and implicitly samples from a distribution shaped by the loss and the noise. This viewpoint suggests that learning may be interpreted simultaneously as:

- **Optimization** toward low loss (the drift term), and
- **Sampling** from an implicit distribution shaped by the noise and the energy landscape—with implications for generalization and flat minima.

5.8 Momentum and Underdamped Langevin

Many optimizers use momentum: the update depends on an accumulated velocity, not just the current gradient. That corresponds to underdamped Langevin, which has a momentum variable. The overdamped Langevin equation (43) is first-order in time: the state is θ_t and there is no explicit momentum. In many optimization settings, however, **momentum** is used (e.g., SGD with momentum, Adam): the update depends not only on the current gradient but on an accumulated velocity. This can be viewed as a discretization of **underdamped** (or **kinetic**) Langevin dynamics. The state is now (θ_t, p_t): position and momentum. The dynamics are second-order, like in mechanics, but with friction and noise:

$$d\theta_t = p_t\,dt, \qquad dp_t = -\nabla\mathcal{E}(\theta_t)\,dt - \gamma p_t\,dt + \sqrt{2\gamma\beta^{-1}}\,dW_t, \tag{53}$$

where p_t is a momentum variable and $\gamma > 0$ is a friction coefficient. In the limit $\gamma \to \infty$, momentum is heavily damped and the system reduces to overdamped Langevin—when friction is very large, momentum is killed quickly and we recover overdamped Langevin. For finite γ, the phase-space flow has a symplectic component (from the Hamiltonian $H = \frac{1}{2}|p|^2 + \mathcal{E}(\theta)$) that is weakly broken by friction and noise. Thus momentum-based optimizers can be understood as sitting between conservative mechanics (symplectic, reversible) and overdamped dissipation (first-order, purely metric): they retain a notion of inertia while still dissipating energy. This perspective links the "weakly broken symplectic structure" question to practical algorithm design.

5.9 Preview: From Thermodynamic Flows to Inference and Generative Modeling

The ideas in this section—Gibbs measure, Fokker–Planck, free energy—will reappear when we discuss distribution-space dynamics, variational inference, and diffusion. The distributional perspective introduced here—equilibrium Gibbs measure, Fokker–Planck evolution, free energy as Lyapunov functional—sets the stage for the next sections:

- **Distribution-space dynamics** (Section 6): we treat ρ_t as the primary state and define "gradient flow" on the space of measures, with Wasserstein and KL geometries.
- **Variational inference**: minimizing $\mathrm{KL}(q\|p)$ is free-energy minimization when p is a posterior; the approximating q is driven by the same thermodynamic tradeoff.
- **Diffusion and score-based generative modeling**: forward diffusion is an entropy-increasing Fokker–Planck flow; the learned reverse process is a controlled drift that inverts it.
- **Schrödinger bridges**: path-space variational principles that interpolate between distributions with entropy-regularized transport.

6 Distribution-Space Dynamics

In the previous section we saw that stochastic dynamics induce an *evolving distribution* ρ_t over parameters, governed by the Fokker–Planck equation. We now take that idea one step further: we treat the *distribution* as the primary state and ask how to define "gradient flow" and "steepest

descent" on the space of probability measures. The answer depends on the **geometry** we choose—and different choices lead to different PDEs and different algorithms. This section introduces the two main geometries (Wasserstein and KL/Fisher) and shows how they recover familiar dynamics.

6.1 From Parameter Trajectories to Evolving Laws

We now take the distribution ρ_t as the main object: instead of following one trajectory, we study how the whole probability distribution over parameters evolves. In the stochastic setting, the learning state is naturally described not only by a single trajectory θ_t, but by its **law** ρ_t: the probability measure on Θ such that $\theta_t \sim \rho_t$. That is, $\rho_t(A) = \mathbb{P}(\theta_t \in A)$ for (Borel) sets $A \subseteq \Theta$. This viewpoint is useful in at least three settings:

- **Stochastic optimization:** randomness induces an evolving ensemble over parameters; we care about the distribution of iterates, not only a single run.
- **Inference:** in Bayesian or variational settings the object of interest is a *posterior distribution* over parameters; we optimize over distributions q that approximate p.
- **Generative modeling:** diffusion and flow-based models explicitly construct a *flow of distributions* from noise to data.

We therefore treat ρ_t as the fundamental state variable and study dynamics directly on the space of probability measures.

Notation. Let $\mathcal{P}(\Theta)$ denote the set of probability measures on Θ (often restricted to measures with finite second moment for technical reasons). So $\mathcal{P}(\Theta)$ is the space of all probability distributions on Θ; a point in this space is a whole distribution, not a single θ. A **distribution-space dynamical system** is a rule that evolves $\rho_t \in \mathcal{P}(\Theta)$ through time—either as a PDE (e.g., Fokker–Planck) or as a discrete update (e.g., a proximal step on measures).

6.2 Transport Viewpoint: The Continuity Equation

Many distributional dynamics can be written as a conservation law: probability mass is neither created nor destroyed, only moved by a velocity field. A large class of distributional dynamics can be written as a **conservation law** (mass transport equation):

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \tag{54}$$

where $v_t(\theta)$ is a **velocity field** on Θ that transports probability mass. This is the same continuity equation that appears in fluid dynamics: ρ_t is like a density of fluid and v_t is the velocity of the fluid. In one dimension, $\partial_t \rho + \partial_x(\rho v) = 0$ simply says that the rate of change of mass in any interval equals the net flux in minus the flux out. Equation (54) expresses that *probability is conserved*: no mass is created or destroyed, only moved around by the flow v_t.

When does diffusion appear? When the underlying dynamics are stochastic (e.g., Langevin), the density evolution includes an additional **diffusion** term, yielding PDEs of Fokker–Planck type:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = \nabla \cdot (D \nabla \rho_t), \tag{55}$$

with diffusion tensor D. The left-hand side is transport by the drift v_t; the right-hand side is spreading due to noise. So when we add noise (Langevin), we get an extra diffusion term in the PDE for ρ_t. The Fokker–Planck equation (45) is of this form with $v_t = -\nabla\mathcal{E}$ and $D = \beta^{-1}I$.

6.3 Functionals on Measures and the Meaning of a Gradient

To define "gradient flow" on the space of measures, we need two things: a functional we want to decrease (e.g. free energy or KL) and a way to measure "distance" between measures—that is, a geometry. In finite dimensions, gradient flow is defined once we choose an energy $\mathcal{E}$ and a metric g: steepest descent is $\dot{\theta} = -\nabla_g\mathcal{E}$. On the space of probability measures $\mathcal{P}(\Theta)$, the same logic applies, but both ingredients require care.

To define **gradient flow in** $\mathcal{P}(\Theta)$, one must specify:

- An **energy functional** $\mathcal{G} : \mathcal{P}(\Theta) \to \mathbb{R}$. Examples: free energy $\mathcal{F}(\rho)$, KL divergence $\mathrm{KL}(\rho\|p)$, or an expected cost $\int c(\theta)\,\rho(\theta)\,d\theta$.
- A **geometry** (metric) on $\mathcal{P}(\Theta)$ that defines distances between measures and therefore what "steepest descent" means. Unlike $\mathbb{R}^d$, the space of measures is infinite-dimensional and admits many inequivalent metrics.

The gradient of $\mathcal{G}$ at ρ is not a vector in $\mathbb{R}^d$ but a **tangent vector** to $\mathcal{P}(\Theta)$ at ρ—in the transport picture, a velocity field v that moves mass. So the "gradient" of a functional on measures is not a vector in $\mathbb{R}^d$ but a field that tells each point how to move its mass. The rate of change of $\mathcal{G}$ in the direction v is given by the **first variation** $\frac{\delta\mathcal{G}}{\delta\rho}$, and the metric converts this into a preferred direction (the gradient). Different metrics yield different "gradients" and hence different PDEs. Which geometry we choose changes what "steepest descent" means and hence the resulting PDE or algorithm.

Two geometries are particularly fundamental for our purposes:

- **Wasserstein geometry** (optimal transport): distance is measured by the cost of *transporting* mass from one distribution to another. Natural for diffusion, fluid-like flows, and physical transport.
- **Information geometry** (KL / Fisher–Rao) [1]: distance is measured by KL divergence or the Fisher information metric. Natural for inference, variational methods, and mirror-descent-type updates.

6.4 Wasserstein Geometry and Otto's Calculus (Steepest Descent as Transport)

In Wasserstein geometry, the distance between two distributions is the minimum cost of *transporting* mass from one to the other when cost is squared distance. The **2-Wasserstein distance** $W_2(\rho_0, \rho_1)$ between two probability measures [15] is the minimal cost of *transporting* the mass of ρ_0 onto ρ_1 when cost is squared Euclidean distance. Informally: imagine ρ_0 and ρ_1 as piles of sand; W_2^2 is the minimum total squared distance that sand particles must move to reconfigure the first pile into the second. So steepest descent in this geometry moves mass in a coordinated way, like a fluid.

This "earth mover's" interpretation makes Wasserstein geometry natural for dynamics that *move* probability mass through Θ (e.g., diffusion, advection).

In Wasserstein geometry, tangent vectors at ρ are represented by velocity fields v that induce a flow satisfying the continuity equation (54). The **Otto calculus** [12] provides a formal Riemannian structure on $\mathcal{P}(\Theta)$ so that the "gradient" of a functional $\mathcal{G}$ is given by a potential. The potential ψ_t is the functional derivative of $\mathcal{G}$ with respect to ρ; it plays the role of "gradient" in this geometry. Steepest descent of $\mathcal{G}$ in this geometry yields dynamics of the form

$$\partial_t \rho_t + \nabla \cdot (\rho_t \nabla \psi_t) = 0, \tag{56}$$

where the potential ψ_t is determined by the **first variation** (functional derivative) of $\mathcal{G}$:

$$\psi_t(\theta) = \frac{\delta \mathcal{G}}{\delta \rho}(\rho_t)(\theta). \tag{57}$$

Equivalently, the velocity field is

$$v_t(\theta) = -\nabla \psi_t(\theta) = -\nabla \left(\frac{\delta \mathcal{G}}{\delta \rho}(\rho_t)(\theta) \right). \tag{58}$$

Interpretation. Wasserstein gradient flow is steepest descent where "distance" between nearby measures is measured by transport cost. So the flow moves mass in the direction that decreases $\mathcal{G}$ most efficiently *per unit of mass moved*. The continuity equation ensures that this motion conserves probability.

6.5 Variational Time Discretization: The JKO Scheme

Just as minimizing movements in parameter space gave gradient flow, a minimizing-movement scheme on measures with Wasserstein penalty gives the JKO scheme. A canonical **variational** characterization of Wasserstein gradient flows is the Jordan–Kinderlehrer–Otto (JKO) minimizing movement scheme [2, 7]:

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \mathcal{G}(\rho) + \frac{1}{2\eta} W_2^2(\rho, \rho_k) \right\}. \tag{59}$$

At each time step we choose the next distribution ρ_{k+1} by minimizing $\mathcal{G}(\rho)$ (the "energy") plus $\frac{1}{2\eta} W_2^2(\rho, \rho_k)$ (a penalty for moving far from ρ_k in Wasserstein distance). So at each step we minimize energy plus a penalty for moving far from the current distribution, where "far" is measured by W_2. This is the distributional analogue of **proximal gradient descent** in parameter space: instead of "parameter + Euclidean penalty," we have "measure + Wasserstein penalty." As $\eta \to 0$, the discrete iterates converge to the continuous-time Wasserstein gradient flow of $\mathcal{G}$.

6.6 Example: Fokker–Planck as Wasserstein Gradient Flow of Free Energy

We can now tie Section 5 to Section 6: the Fokker–Planck equation is exactly the Wasserstein gradient flow of the free energy $\mathcal{F}$. A central example links directly to Section 5. Consider the free energy functional

$$\mathcal{F}(\rho) = \int \mathcal{E}(\theta) \rho(\theta)\, d\theta + \beta^{-1} \int \rho(\theta) \log \rho(\theta)\, d\theta. \tag{60}$$

The first variation of $\mathcal{F}$ has two parts: one from $\int \mathcal{E}\rho$ and one from the entropy term; together they give $\frac{\delta \mathcal{F}}{\delta \rho} = \mathcal{E} + \beta^{-1} \log \rho + \text{const}$. In Otto's calculus, the Wasserstein gradient flow of $\mathcal{F}$ is therefore given by $v_t = -\nabla(\mathcal{E} + \beta^{-1} \log \rho_t)$, which when substituted into the continuity equation and combined with the diffusion that arises from the $\log \rho$ term yields precisely the Fokker–Planck equation:

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \mathcal{E}) + \beta^{-1} \Delta \rho_t. \tag{61}$$

So the drift term comes from the energy and the Laplacian from the entropy; thermodynamic relaxation is steepest descent of $\mathcal{F}$ in Wasserstein geometry. This provides a variational and geometric explanation for thermodynamic relaxation: the system evolves by following the direction of greatest free-energy decrease, with distance measured by transport cost.

6.7 Information Geometry: KL and Fisher–Rao Flows

For inference and variational methods, a different geometry is often used: distance is measured by KL divergence or the Fisher–Rao metric, not by transport cost. Instead of measuring distance by transport cost, we measure it by **KL divergence** $\mathrm{KL}(\rho \| \rho_k) = \int \rho \log(\rho/\rho_k)\, d\theta$ or by the **Fisher–Rao** (information) metric. A prototypical variational update takes the form

$$\rho_{k+1} = \arg\min_{\rho} \left\{ \mathcal{G}(\rho) + \frac{1}{\eta} \mathrm{KL}(\rho \| \rho_k) \right\}, \tag{62}$$

which is the distribution-space analogue of **mirror descent**: we minimize the functional plus a KL penalty for deviating from the current ρ_k. In parametric families $\rho = \rho_\phi$ (e.g., Gaussians parameterized by mean and covariance), this KL-based geometry leads to **natural gradient** updates, where the gradient is preconditioned by the Fisher information matrix.

When to use which geometry?

- **Wasserstein:** evolution is thought of as *transporting mass* through Θ (particles move; diffusion; fluid-like flows). Natural for Langevin, Fokker–Planck, and physical transport.
- **KL / Fisher:** evolution is thought of as *reweighting or deforming the density* (mass does not "travel" in the same sense). Natural for variational inference, Bayesian updates, and mirror descent.

Roughly: Wasserstein when we think of moving mass; KL/Fisher when we think of reweighting or deforming the density. The same functional $\mathcal{G}$ can have different gradient flows depending on which metric we choose; the PDE and the discrete algorithm both change.

6.8 Particle Systems and Mean-Field Limits

Distributional dynamics often arise as the limit of many particles: if we run N copies of a stochastic process and look at the empirical distribution, as $N \to \infty$ it converges to a deterministic flow satisfying a PDE. Consider N particles $\{\theta_t^{(i)}\}_{i=1}^N$, each evolving (e.g., by independent Langevin dynamics or by dynamics that depend on the empirical distribution). The **empirical measure** is the discrete distribution that puts mass $1/N$ at each particle:

$$\rho_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_t^{(i)}}. \tag{63}$$

So ρ_t^N is the histogram of the N particles; as N grows, this histogram behaves like a smooth density obeying Fokker–Planck or a similar equation. As $N \to \infty$, under suitable conditions ρ_t^N converges (in the sense of weak convergence of measures) to a deterministic measure ρ_t whose density satisfies a PDE such as (61) or more general **McKean–Vlasov** equations. Intuition: a single particle is random, but the *distribution* of a large population becomes deterministic and obeys a continuum law. This provides a bridge between *microscopic* learning dynamics (individual parameter trajectories, e.g., many SGD runs) and *macroscopic* evolution (the flow of ρ_t).

6.9 Preview: Inference and Generative Modeling as Distribution Flows

The next section will show that variational inference and diffusion models are special cases of distribution-space flows with specific choices of functional and geometry. The distribution-space viewpoint provides the natural language for Section 7:

- **Variational inference:** we optimize over distributions q that approximate a target p; the objective is KL or free energy, and the geometry is often KL/Fisher.
- **Diffusion models:** the forward process is a distribution flow (Fokker–Planck); the learned reverse process is a controlled flow that maps noise to data.
- **Schrödinger bridges:** path-space variational principles that interpolate between two given marginals with entropy-regularized transport.

7 Connections to Optimization, Variational Inference, and Diffusion

The preceding sections developed a hierarchy of dynamical and variational structures:

- **Section 4:** deterministic gradient flows on parameter space (energy dissipation; minimizing movements in Euclidean geometry).
- **Section 5:** stochastic Langevin dynamics with thermodynamic structure (Gibbs equilibrium; Fokker–Planck; free energy as Lyapunov functional).
- **Section 6:** gradient flows of functionals on the space of probability measures (Wasserstein vs KL geometry; JKO scheme).

We now show that several central paradigms in modern machine learning arise as **specializations** of this general framework. In each case we identify: the *state variable* (parameters vs measures), the *functional* being minimized, the *geometry* that defines steepest descent, and the *stochastic structure* (if any). This unified view clarifies how optimization, inference, and generative modeling relate and suggests that algorithm design can be understood as choosing a variational flow with prescribed geometric and thermodynamic properties.

7.1 Deterministic Optimization as Metric Gradient Flow

We start by placing classical gradient descent in our framework: it is gradient flow of the loss in Euclidean (or Riemannian) geometry, with no noise.

State: parameter $\theta \in \Theta$. **Functional:** $\mathcal{E}(\theta)$ (loss or risk). **Geometry:** Euclidean (or a Riemannian metric g). **Stochasticity:** none.

Classical gradient descent is the continuous-time limit of the discrete update $\theta_{k+1} = \theta_k - \eta \nabla \mathcal{E}(\theta_k)$:

$$\dot{\theta} = -\nabla \mathcal{E}(\theta). \tag{64}$$

This is precisely the gradient flow of $\mathcal{E}$ in Euclidean geometry (Section 4). Energy decreases monotonically: $\frac{d}{dt}\mathcal{E}(\theta(t)) = -\|\nabla \mathcal{E}\|^2 \leq 0$.

When the geometry is changed, the *direction* of steepest descent changes:

- **Natural gradient descent** [1] uses the Fisher information metric (the Riemannian metric induced by the statistical model). It is often better suited to parameter spaces with curvature (e.g., distributions) and can yield faster convergence in certain regimes.
- **Mirror descent** [11] corresponds to proximal updates under a Bregman divergence; the geometry is determined by a convex potential. It generalizes gradient descent to non-Euclidean geometry and is closely related to KL-based updates in distribution space.

So "choosing an optimizer" can be read as choosing a metric on parameter space. All such methods may be viewed as gradient flows $\dot{\theta} = -\nabla_g \mathcal{E}(\theta)$ for a chosen Riemannian metric g. Thus **optimizer design is geometry choice**: we are selecting how to measure "distance" and "steepest" on parameter space.

7.2 Stochastic Optimization and Implicit Sampling

SGD adds noise to the gradient; under suitable scaling that noise can be approximated by Langevin dynamics, so SGD both optimizes and implicitly samples.

State: parameter θ (and implicitly its law ρ_t). **Functional:** $\mathcal{E}(\theta)$; at equilibrium, the law is Gibbs-like. **Geometry:** Euclidean on Θ; distribution flow has Wasserstein structure. **Stochasticity:** gradient noise (minibatch).

Stochastic gradient descent introduces noise through minibatch sampling:

$$\theta_{k+1} = \theta_k - \eta \widehat{\nabla \mathcal{E}}(\theta_k). \tag{65}$$

Under appropriate scaling limits (small step size, large minibatch), the rescaled process can often be approximated by an SDE of Langevin type:

$$d\theta_t = -\nabla \mathcal{E}(\theta_t)\, dt + (\eta \Sigma(\theta_t))^{1/2} dW_t. \tag{66}$$

where Σ encodes the covariance of the gradient noise.

From the thermodynamic perspective of Section 5, such dynamics do two things at once: (i) they **optimize** (drift toward low $\mathcal{E}$), and (ii) they **sample** from an implicit equilibrium distribution

concentrated near low-energy regions, with spread controlled by the noise scale. So the noise structure of SGD shapes which solutions we tend to find. SGD does not simply converge to a single minimizer; it explores a region of parameter space, and the noise structure (e.g., batch size, learning rate) shapes the **implicit bias** and generalization of the learned solution. This dual interpretation links optimization to sampling and has been used to explain why SGD often finds flatter minima that generalize well.

7.3 Variational Inference as Free Energy Minimization

Variational inference approximates a target distribution p (e.g. a posterior) by minimizing KL divergence from an approximating family to p. When p is Gibbs, that is exactly free-energy minimization.

State: distribution $q \in \mathcal{P}(\Theta)$ (the variational approximation). **Functional:** $\mathrm{KL}(q\|p)$ or equivalently free energy $\mathcal{F}(q)$. **Geometry:** typically KL or Fisher–Rao. **Stochasticity:** optional (e.g., reparameterized gradients).

Variational inference (VI) [4] replaces optimization over a single parameter with **optimization over distributions**. We have a target distribution $p(\theta)$ (e.g., a Bayesian posterior) that is intractable to sample from or normalize. We choose an approximating family $\{q_\phi\}$ and select q by minimizing

$$\mathrm{KL}(q\|p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta)}\, d\theta. \tag{67}$$

We use the reverse KL $\mathrm{KL}(q\|p)$ so that q is encouraged to put mass where p has mass and to avoid placing mass where p is tiny.

When the target is a Gibbs distribution, $p(\theta) \propto e^{-\beta \mathcal{E}(\theta)}$ (e.g., posterior $\propto$ likelihood $\times$ prior with $\mathcal{E} = -\log(\text{likelihood} \cdot \text{prior})$), we have

$$\mathrm{KL}(q\|p) = \beta \int \mathcal{E}(\theta) q(\theta)\, d\theta + \int q(\theta) \log q(\theta)\, d\theta + \text{const.} \tag{68}$$

Up to constants and the factor β^{-1}, this is precisely the free energy functional

$$\mathcal{F}(q) = \int \mathcal{E}(\theta) q(\theta)\, d\theta + \beta^{-1} \int q(\theta) \log q(\theta)\, d\theta. \tag{69}$$

Thus **variational inference is minimization of free energy in distribution space**. The variational posterior q is the best approximation to p in the sense of KL, and the same tradeoff (expected energy vs entropy) that governs Langevin equilibrium appears here. When combined with KL-based proximal geometry (Section 6), VI updates can be seen as mirror-descent-type flows on $\mathcal{P}(\Theta)$. Practical algorithms (e.g., reparameterization, natural gradients) implement discrete approximations to these flows.

Example (mean-field Gaussian VI). A concrete instance is variational inference with a Gaussian approximating family $q_\phi(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$, where $\phi = (\mu, \Sigma)$ (or a parameterization of Σ, e.g. a Cholesky factor). The objective $\mathrm{KL}(q_\phi\|p)$ is minimized over ϕ. The natural gradient of this objective with respect to ϕ is preconditioned by the Fisher information matrix of the Gaussian family; the resulting flow on (μ, Σ) is a discrete approximation to the KL-gradient flow of free energy on $\mathcal{P}(\Theta)$. So the usual mean-field Gaussian VI update is steepest descent in Fisher–Rao geometry on the space of Gaussians.

7.4 Diffusion Models as Controlled Distribution Flows

Diffusion models have two phases: a forward process that turns data into noise (a distribution flow that increases entropy) and a learned reverse process that turns noise back into data.

State: distribution over data (or latent) space; we explicitly model ρ_t along a path from data to noise and back. **Functional:** in the forward direction, entropy increase; in the reverse, a learned drift. **Geometry:** Wasserstein / transport. **Stochasticity:** central (forward and reverse SDEs).

Score-based and diffusion generative models [6, 14] construct probability flows in two phases.

Forward diffusion is a stochastic process that gradually adds noise to data:

$$dx_t = f(x_t, t)\, dt + g(t)\, dW_t. \tag{70}$$

Typically $x_0 \sim p_{\text{data}}$ and as t increases the distribution of x_t approaches a simple noise distribution (e.g., Gaussian). The associated density evolution satisfies a Fokker–Planck equation; the process **increases entropy** and "dissolves" the data distribution into noise.

Reverse process. Under suitable conditions, the time-reversal of the forward SDE is again an SDE with a modified drift that depends on the **score** $\nabla \log \rho_t(x)$. The score tells us the direction in which the density increases fastest; the reverse SDE uses it to push mass from noise toward data. Generative sampling is then: start from noise and run the reverse SDE, using a learned (e.g., neural) approximation to the score. The learned drift effectively **inverts** the forward diffusion and pushes the noise distribution back toward the data distribution.

So diffusion is a controlled flow on the space of distributions: forward is Fokker–Planck; reverse is learned transport. From the distribution-space viewpoint, diffusion models are **controlled distribution flows**: the forward flow is a Fokker–Planck (Wasserstein gradient flow of free energy); the reverse flow is a learned transport that undoes it. They can also be viewed as entropy-regularized optimal transport between the noise and data distributions, or as approximations to Schrödinger bridges.

7.5 Schrödinger Bridges and Entropic Interpolation

The **Schrödinger bridge** problem [10] asks: given two distributions at two times, what is the most likely (or entropy-regularized) stochastic process that connects them? Equivalently, given two probability distributions ρ_0 and ρ_T at times 0 and T, what is the *most likely* stochastic process (or the one that minimizes relative entropy to a reference process) that has these marginals? It can be formulated as

$$\min_{\mathbb{P}} \mathrm{KL}(\mathbb{P} \| \mathbb{P}_0) \tag{71}$$

over path measures $\mathbb{P}$ subject to $\mathbb{P}$ having marginals ρ_0 and ρ_T at the endpoints; $\mathbb{P}_0$ is a reference (e.g., Brownian) bridge. The solution is an entropy-regularized interpolation between the two marginals.

This framework unifies:

- **Optimal transport:** the limit of zero noise gives deterministic mass transport (Monge problem).

- **Diffusion processes:** the solution is a diffusion with a drift that enforces the endpoint constraints.
- **Entropy-regularized variational principles:** the objective balances fidelity to the reference process with the marginal constraints.

Diffusion models can be seen as learning an approximation to such a Schrödinger bridge from noise to data: the forward process is the reference, and the learned score defines the reverse process that (approximately) interpolates from noise to data.

7.6 Structural Summary

The following table summarizes how each method fits into the same variational–geometric picture; the differences are state variable, functional, and geometry. We may summarize the structural relationships as follows. Each paradigm is a variational flow; what changes is the state variable, the functional, and the geometry.

Method	**State**	**Functional**	**Geometry / Structure**
Gradient Descent	θ	$\mathcal{E}(\theta)$	Euclidean metric
Natural Gradient	θ	$\mathcal{E}(\theta)$	Fisher metric
Langevin / SGD	θ, law ρ_t	$\mathcal{E}$; equil. Gibbs	Euclidean + noise; W_2 for ρ
Variational Inference	$q \in \mathcal{P}(\Theta)$	KL / Free energy	KL / Fisher geometry
Diffusion Models	ρ_t (data space)	Forward: entropy $\uparrow$; reverse: learned	Stochastic transport

So algorithm design can be viewed as choosing these structural ingredients. These paradigms differ not in their core objective of functional optimization, but in the *state variable* (parameter vs distribution), the *functional* (energy vs free energy vs KL), and the *geometric and stochastic structures* imposed on the state space. Recognizing this unity helps when designing or analyzing new algorithms: we can ask which geometry and which functional they implicitly use.

7.7 Interpretive Perspective

We can summarize the essay's viewpoint in one sentence. Modern learning algorithms may therefore be viewed as instances of a broader principle:

> **Learning is variational flow under a chosen geometry and stochastic structure.**

That is, each algorithm can be understood as (possibly the discrete approximation to) a dynamical system that decreases a certain functional along a direction of "steepest descent," where steepest is defined by a metric on the state space. Different algorithmic families correspond to different choices of:

- **State variable:** parameters θ (optimization, SGD) vs. distributions ρ or q (VI, diffusion).

- **Functional:** energy $\mathcal{E}$, free energy $\mathcal{F}$, KL divergence, or an entropy-regularized transport cost.
- **Geometry:** Euclidean, Riemannian (e.g., Fisher), Wasserstein (transport), or KL/Fisher on measures.
- **Noise structure:** deterministic (gradient descent) vs. stochastic (Langevin, SGD, diffusion).

This does not say that learning *is* mechanics, but that both can be described in the same variational language with different choices of geometry and dissipation. This perspective does not reduce learning to mechanics—conservative and dissipative dynamics remain distinct—but it organizes a wide range of methods under a single variational–geometric language and suggests that new algorithms can be designed by explicitly choosing these structural ingredients.

7.8 Closing Reflections

To close, we restate the main message and its implications for future work. This note developed a variational–geometric lens in which the qualitative behavior of learning dynamics is determined less by the specific algorithmic surface form than by a small set of structural choices: the *state variable* (parameters versus distributions), the *functional* being optimized (energy, free energy, KL), the *geometry* used to define steepest descent (Euclidean/Riemannian, KL/Fisher, Wasserstein), and the *stochastic structure* that injects fluctuations and induces thermodynamic relaxation. From this viewpoint, conservative mechanics and dissipative learning are not conflated; rather, they occupy distinct corners of a common variational landscape, with symplectic structure enforcing reversibility and metric structure enforcing contraction. Stochasticity then lifts parameter trajectories to distribution flows, where free energy becomes a Lyapunov functional and diffusion admits a clean variational characterization. The unifying message is therefore structural: optimization, variational inference, and diffusion can be read as different instantiations of variational flow under different geometries and noise models. This perspective suggests concrete research directions—most notably, designing learning dynamics by explicitly choosing (or interpolating between) geometric invariants, dissipation mechanisms, and fluctuation scales—and motivates treating "optimizer design" as the design of a dynamical system with prescribed variational and thermodynamic properties.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2nd edition, 2008.

[3] Vladimir I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 2nd edition, 1989.

[4] 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.

[5] Herbert Goldstein, Charles Poole, and John Safko. *Classical Mechanics*. Addison-Wesley, 3rd edition, 2001.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[7] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[8] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[9] Paul Langevin. Sur la théorie du mouvement brownien. *Comptes Rendus de l'Académie des Sciences*, 146:530–533, 1908.

[10] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems - Series A*, 34(4):1533–1574, 2014.

[11] Arkadii Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[12] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[13] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*, volume 18 of *Springer Series in Synergetics*. Springer, 2nd edition, 1996.

[14] Yang Song, Chen Meng, and Stefano Ermon. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[15] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[16] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011.