Learning Dynamics

Eric Hermosis
eric.hermosis@gmail.com

January 7, 2026

Abstract

The work presents a fundamental theory of artificial learning that aims to explain existing optimization algorithms used to train artificial intelligence models and to support the development of new ones. It is based on a thermodynamic framework, employing geometric and a Hamiltonian formalism to derive the evolution equations of learning. These equations provide a physical grounding for the learning process and show that it can be understood as a thermodynamic process in which models evolve according to the information they perceive.

Introduction

This work aims to develop a fundamental theory of artificial learning that explains existing optimization algorithms, facilitates their improvement, and enables the creation of new ones. To this end, we make only three assumptions about a model capable of learning:

- It can be described by a finite set of parameters.
- It is differentiable with respect to its parameters.
- It makes no assumptions about the information it ingests.

The theory is constructed based on the axiomatic thermodynamic framework proposed by Callen [1]. Assuming a quasi-static regime, symplectic geometry is used to describe the geometry of the model's phase space [2], and a Hamiltonian formalism [3] is employed to derive the model's equations of evolution.

Finally, we use these equations to re-derive established optimization algorithms, thereby validating the theory and offering a physical grounding for the learning process. This shows that algorithms like momentum-based Stochastic Gradient Descent [4] or regularization techniques like Weight Decay [5] emerge naturally as consequences of the proposed equations of learning.

Model

The connection between the concepts of information and entropy [6] suggests that learning can be modeled as a thermodynamic process, where the participants, known as models, evolve based on the information they perceive.

A model is a simplified representation of a system, defined by a set of parameters that determine its behavior. A specific choice of a set of parameters defines what we will call a configuration.

We say that a model is differentiable if its configurations are points on a smooth differential manifold, allowing learning to be defined in terms of a parametrized curve over manifold. [7]

On the other hand, we will only focus on models that are domain-agnostic, meaning they do not incorporate assumptions about the information they will ingest. In this way, the dynamics of their learning can be designed solely in terms of their configurations.

Equilibrium

Let's consider a differentiable model whose parameters reside in a smooth manifold S. For each dimension of the manifold, we can associate a parameter so that each configuration $s \in \mathcal{S}$ can be described in terms of coordinates.

$$(U/c, w^1, ..., w^d) = (w^0, \mathbf{w}) \in \mathbb{R}^{d+1} \tag{1}$$

Where U represents the internal energy, $w^1, \ldots, w^d$ are extensive parameters of the model known as weights, and c is a constant such that w^0 is dimensionless.

We say that the model is in an equilibrium state [1] if an entropy function S can be defined over it, measured in units of information, and monotonically increasing with respect to energy, that is:

$$\frac{\partial S}{\partial U} > 0 \tag{2}$$

By differentiating S, we can see how it changes under infinitesimal displacements of the configuration:

$$dS = \frac{\partial S}{\partial U} dU + \sum_j \frac{\partial S}{\partial w^j} dw^j \tag{3}$$

The rates of change of entropy along the directions of energy and weights give rise to the conjugate variables:

$$\beta = \frac{\partial S}{\partial U} \qquad Y_j = \frac{\partial S}{\partial w^j} \qquad j = 1, ..., d \tag{4}$$

These variables are known as the intensive parameters of the model. We will refer to the $\mathbf{Y}$ intensive parameters as entropic moments. We can also identify the temperature T as the reciprocal of the parameter β conjugate to the energy, that is:

$$T \equiv \frac{1}{\beta} > 0 \tag{5}$$

The entropy function is local, that is, it is only defined for each equilibrium state. Therefore, if one seeks to describe the states of the model over the entire state space, it is necessary to resort to the phase space [3] defined over $\mathcal{S}$.

The phase space is a construction over the state space that assigns to each point its cotangent space; that is, if $(w_0, \mathbf{w})$ are coordinates of the state space, then $(w_0, \mathbf{w}, Y^0, \mathbf{Y})$ are coordinates of the phase space. Let us now consider the 1-form living in the phase space Ω given by:

$$\omega = \beta dU + \sum_j Y_j dw^j \in \Omega \tag{6}$$

This form generalizes the notion of the differential of entropy, such that the model is in an equilibrium state if there exists an entropy function S such that:

$$\omega = dS \tag{7}$$

Expanding the exterior derivative of the differential 1-form ω, we obtain the differential 2-form:

$$d\omega = d\beta \wedge dU + \sum_j dY_j \wedge dw^j \tag{8}$$

The latter is known as the symplectic form [2] and allows the phase space Ω to be endowed with a Hamiltonian geometric structure.

Evolution

For each point in the phase space, entropy is defined only for equilibrium states. This means that, to remain within the scope of a thermodynamic description, the system's evolution must be slow enough to preserve the quasi-static approximation, then the learning curve can be viewed as a succession of equilibrium states what allows us to define canonical pairs over the entire manifold through Poisson brackets:

$$\{U, \beta\} = 1 \qquad \{w^i, Y_j\} = \delta^i_j \qquad \text{with } \delta^i_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{9}$$

Then, we can recover an analogue to Hamilton's equations [8] for thermodynamic parameters describing the evolution of the model parameters without constraints:

$$-h\frac{dw^i}{dt} = k\{w^i, H\} = k\frac{\partial H}{\partial Y_i} \qquad -\frac{dY_i}{dt} = k\{Y_i, H\} = -k\frac{\partial H}{\partial w^i} \tag{10}$$

$$-\frac{dU}{dt} = k\{U, H\} = k\frac{\partial H}{\partial \beta} \qquad -h\frac{d\beta}{dt} = k\{\beta, H\} = -k\frac{\partial H}{\partial U} \tag{11}$$

Where h is the unit of action and k is the unit of information, which are introduced to maintain consistent units.

The problem with this formulation is that it leads to dynamics in which the model evolves in closed orbits. To address this, we introduce a coupling of the intensive parameters with the temperature of the form:

$$Y_i = \beta X_i \qquad i = 1, ..., d \tag{12}$$

We will refer to the $\mathbf{X}$ parameters as energy moments. This coupling is not arbitrary, rather, it arises directly from the energy representation of thermodynamics:

$$dU = TdS - \sum_j X_j dw^j \tag{13}$$

The coupling deforms the symplectic structure that describes the geometry of the phase space. By substituting the coupling into the 2-form $d\omega$, we obtain:

$$d\omega = d\beta \wedge (dU + \sum_j X_j dw^j) + \beta \sum_j dX_j \wedge dw^j \tag{14}$$

Which remains a non-degenerate symplectic form for $\beta > 0$, a condition that has already been imposed. The new non-negative Poisson brackets yield:

$$\{U, \beta\} = 1 \qquad \{w^i, X_j\} = \delta^i_j \qquad \{U, X_i\} = -\frac{1}{\beta}X_i \tag{15}$$

And their respective equations of motion are given by:

$$-h\frac{dw^i}{dt} = \frac{k}{\beta}\frac{\partial H}{\partial X_i} \qquad -h\frac{dX_i}{dt} = -\frac{k}{\beta}\frac{\partial H}{\partial w^i} + \frac{kX_i}{\beta}\frac{\partial H}{\partial U} \tag{16}$$

$$-h\frac{dU}{dt} = k\frac{\partial H}{\partial \beta} - \frac{k}{\beta}\sum_j X_j \frac{\partial H}{\partial X_j} \qquad -h\frac{d\beta}{dt} = -k\frac{\partial H}{\partial U} \tag{17}$$

We will refer to these as the Hermosis equations of learning. While they can be rigorously derived from the new symplectic form, a more streamlined derivation based on the properties of Poisson brackets is provided in the Appendix.

Integration

The presented four equations allow us to describe the learning process of a model. Since they need to be numerically integrated to perform optimization, we seek their integral form. By substituting the last equation into the second one and rearranging the terms, we obtain the evolution equation for the momentum X_i as

$$h\beta\frac{dX_i}{dt} + h\frac{d\beta}{dt}X_i = h\frac{d}{dt}(\beta X_i) = k\frac{\partial H}{\partial w^i} \tag{18}$$

Integrating over the interval $[t-\tau, t]$ with τ small enough to preserve quasi-static aproximation, we obtain an update rule for the momenta:

$$\beta(t)\mathbf{X}(t) = \beta(t-\tau)\mathbf{X}(t-\tau) - \int_{t-\tau}^{t}\mathbf{F}(t')dt' \tag{19}$$

Where $\mathbf{F}$ represents a generalized force, whose components are given by:

$$F_i = -\frac{k}{h}\frac{\partial H}{\partial w^i} \tag{20}$$

On the other hand, integrating the first equation over the same interval, we obtain an update rule for the weights:

$$\mathbf{w}(t) = \mathbf{w}(t-\tau) - \int_{t-\tau}^{t}\mathbf{v}(t')dt' \tag{21}$$

Where $\mathbf{v}$ denotes the learning velocity, with components:

$$v^i = \frac{k}{h\beta}\frac{\partial H}{\partial X_i} \tag{22}$$

This last equation tells us something important, the β parameter determines the system's inertia throughout its evolution. Large values of β slow down learning, while small values accelerate it.

Application

Let us now examine the connection between the proposed dynamics and the current algorithms used in machine learning.

In practice, a model is trained by minimizing a loss function L, which measures the distance between a model's current state and an expected state. Drawing from classical mechanics, we propose a potential analogous to the gravitational potential:

$$V = \frac{\beta c^2}{k}L \tag{23}$$

Where c^2 with units of square energy is introduced just to ensure that L remains dimensionless. The choice of this potential is not arbitrary, it is based on the interpretation of the term β/k as a thermal mass that amplifies the importance of the distance within the potential energy. Furthermore, we propose a kinetic energy in terms of a mass tensor M_{ij} of the form:

$$K = \frac{1}{2}\sum_{ij}M^{ij}Y_iY_j = \frac{\beta}{2k}\sum_{ij}g^{ij}X_iX_j = \frac{\beta}{2k}\mathbf{X}^2 \tag{24}$$

Where g_{ij} is a dimensionless metric tensor, which we assume for now to be constant. In this way, the Hamiltonian is defined as:

$$H = \frac{\beta}{2k}\mathbf{X}^2 + \frac{\beta c^2}{k}L + E(U,\beta) \tag{25}$$

Here E is a function of the temperature, enabling us to specify an arbitrary thermal profile. Under this Hamiltonian, the velocity components are given by:

$$v^i = \frac{k}{h\beta}\frac{\partial H}{\partial X_i} = \frac{1}{h}X^i \qquad X^i = \sum_{ij} g^{ij}X_j \tag{26}$$

Applying an Euler discretization, the weight update rule can be approximated as:

$$\mathbf{w}(t) \approx \mathbf{w}(t-\tau) - \frac{\tau}{h}\mathbf{X}(t) \tag{27}$$

On the other hand, the generalized force driving the learning process has the following components:

$$F_i = -\frac{\beta c^2}{h}\frac{\partial L}{\partial w^i} \tag{28}$$

By evaluating the components and performing the integration, we arrive at an expression for the impulse as a function of the loss gradient.

$$\mathbf{I}(t) = -\int_{t-\tau}^{t}\mathbf{F}(t')dt' = \frac{c^2}{h}\nabla L\int_{t-\tau}^{t}\beta(t')dt' \tag{29}$$

Substituting into the momenta update rule, we obtain:

$$\mathbf{X}(t) = \frac{\beta(t-\tau)}{\beta(t)}\mathbf{X}(t-\tau) + \frac{c^2}{h}\left(\frac{1}{\beta(t)}\int_{t-\tau}^{t}\beta(t')dt'\right)\nabla L \tag{30}$$

This tells us that the momenta are updated in terms of time averages of the thermal mass.

Constant temperature

Assuming a constant temperature, that is, a constant thermal mass:

$$\beta(t) = \beta \tag{31}$$

We obtain a constant impulse that depends on the step size:

$$\mathbf{I}(t) = \frac{\beta c^2}{h}\nabla L\int_{t-\tau}^{t}dt' = \frac{\tau c^2}{h}\beta\nabla L \tag{32}$$

And the update rule will be a historical accumulation of the loss gradient:

$$\mathbf{X}(t) = \mathbf{X}(t-\tau) + \frac{\tau c^2}{h}\nabla L \tag{33}$$

Under a suitable reparameterization, we recover the stochastic gradient descent (SGD) update rule [4]:

$$\begin{aligned}\mathbf{X}(t) &= \mathbf{X}(t-\tau) + \zeta\nabla L \\ \mathbf{w}(t) &\approx \mathbf{w}(t-\tau) - \eta\mathbf{X}(t)\end{aligned} \tag{34}$$

Where η is the learning rate and ζ controls the influence of the loss gradient on the momentum.

Harmonic potential

By adding a harmonic potential to the Hamiltonian such that:

$$H = \frac{\beta}{2k}\mathbf{X}^2 + \frac{\lambda}{2}\mathbf{w}^2 + V \tag{35}$$

The generalized force now includes a term consistent with that used in gradient descent with weight decay [5], which proposes adding a term $\lambda\mathbf{w}$ to the loss gradient. This is because an extra term is now added to the force:

$$F_i' = \lambda w_i \tag{36}$$

Further physical explanations of the underlying mechanism, as well as possible corrections, are deferred to future work.

Exponential cooling

Assuming that the system temperature decreases to zero exponentially, that is, an exponentially growing thermal mass,

$$\beta(t) = \beta e^{\gamma t} \tag{37}$$

We obtain an impulse given by:

$$\mathbf{I}(t) = \frac{\beta c^2}{h}\nabla L \int_{t-\tau}^{t} e^{\gamma t'} dt' = \frac{c^2}{h}\frac{1 - e^{-\gamma\tau}}{\gamma}\beta e^{\gamma t} \tag{38}$$

Then, momentum will be an exponential moving average [9] of the loss gradient.

$$\mathbf{X}(t) = e^{-\gamma\tau}\mathbf{X}(t-\tau) + \frac{1 - e^{-\gamma\tau}}{\gamma}\frac{c^2}{h}\nabla L \tag{39}$$

The exponential moving average is biased toward zero at early times due to its initialization, therefore, a bias correction factor should be applied [10]:

$$\hat{\mathbf{X}}(t) = \frac{\mathbf{X}(t)}{1 - e^{-\gamma t}} \tag{40}$$

Under reparameterization, this recover gradient descent with momentum and friction [11]:

$$\begin{aligned} \mathbf{X}(t) &= \mu\mathbf{X}(t-\tau) + (1-\mu)\zeta\nabla L \\ \mathbf{w}(t) &= \mathbf{w}(t-\tau) - \eta\mathbf{X}(t) \end{aligned} \tag{41}$$

Most implementations of SGD with momentum ignore the γ parameter and omit the bias correction.

Relativistic Hamiltonian

Lastly, we adopt a classical Hamiltonian incorporating relativistic kinetic energy, given by:

$$H = \sqrt{\mathbf{X}^2 + \frac{k^2}{\beta^2}} + V + E(U, \beta) \tag{42}$$

This Hamiltonian is a generalization of the previous one, since for small momenta $|\mathbf{X}| \ll k/\beta$, the kinetic energy can be approximated as:

$$\sqrt{\mathbf{X}^2 + \frac{k^2}{\beta^2}} = \frac{k^2}{\beta^2}\sqrt{\frac{\beta^2}{k^2}\mathbf{X}^2 + 1} \approx \frac{k}{\beta} + \frac{\beta}{2k}\mathbf{X}^2 - \frac{\beta^3(\mathbf{X}^2)^2}{8k^3} + \cdots \tag{43}$$

The first term, kT, corresponds to a rest energy, while higher-order terms are typically discarded. However, recent work on physics-inspired optimizers [12] has shown that retaining higher-order terms in the series can enhance optimization. Under a relativistic regime, the learning velocity will then be:

$$v^i = \frac{k}{h\beta}\frac{\partial H}{\partial X_i} = \frac{k}{h\beta}\frac{X^i}{\sqrt{\mathbf{X}^2 + k^2/\beta^2}} = \frac{1}{h}\frac{X^i}{\sqrt{\beta^2\mathbf{X}^2/k^2 + 1}} \tag{44}$$

If we consider $\beta(t) = \beta$ to be constant, we recover the relativistic gradient descent [13], which proposes weight updates of the form:

$$\mathbf{w}(t) = \mathbf{w}(t - \tau) - \eta\frac{\mathbf{X}}{\sqrt{\mathbf{X}^2 + k^2/\beta^2}} \tag{45}$$

However, considering $\beta(t) = \beta e^{\gamma t}$, we observe that $\mathbf{v} \to 0$ rapidly and the system ceases to learn. This is not an error in the theory, rather, it is due to the fact that we are working under a classical approximation in which the potential is considered decoupled from the metric tensor.

The work of Guskov and Vanchurin on covariant gradient descent [14] suggests that by embedding the potential into the metric tensor, recover adaptive momentum-based optimizers such as Adam [10] and, as a specific case, the RMSProp optimizer [15]. Verifying this correspondence explicitly within the current theoretical framework remains as future work.

Conclusion

The proposed formalism demonstrated that a Hamiltonian framework can recover the majority of current optimization algorithms. This approach moves beyond purely heuristic updates, offering a clear physical interpretation where:

- The loss function acts as a metric distance between model's configurations.
- Temperature dictates the effective mass, weighting the importance of these distances.
- Phase space convergence is guaranteed by the evolution of internal energy.

These results lay the groundwork for a new class of physically-informed optimizers. By identifying where current classical approximations fail, such as the decoupling of the potential from the metric tensor, we open new avenues for incorporating covariant and relativistic dynamics into machine learning, potentially leading to more stable and faster convergence in complex loss landscapes.

Appendix: Derivation of learning equations

The equations for the system without temperature coupling, described by the variables $\mathbf{Y}$, are:

$$-h\frac{dw^i}{dt} = k\{w^i, H\} = k\frac{\partial H}{\partial Y_i} \qquad -h\frac{dY_i}{dt} = k\{Y_i, H\} = -k\frac{\partial H}{\partial w^i} \tag{46}$$

$$-h\frac{dU}{dt} = k\frac{\partial H}{\partial \beta} \qquad -h\frac{d\beta}{dt} = -k\frac{\partial H}{\partial U} \tag{47}$$

Let now derive the equations with the coupling $\mathbf{Y} = \beta\mathbf{X}$. Using the Leibniz rule for the bracket containing the coordinates of $\mathbf{Y}$, for a Hamiltonian H, we have:

$$\{Y_i, H\} = \{\beta X_i, H\} = \beta\{X_i, H\} + X_i\{\beta, H\} \tag{48}$$

Given that:

$$\{Y_i, H\} = -\frac{\partial H}{\partial w^i} \qquad \{\beta, H\} = -\frac{\partial H}{\partial U} \tag{49}$$

We can substitute these values to obtain the Poisson brackets of the system in terms of the momenta $\mathbf{X}$:

$$-\frac{\partial H}{\partial w^i} = \beta\{X_i, H\} - X_i\frac{\partial H}{\partial U} \Rightarrow \{X_i, H\} = -\frac{1}{\beta}\frac{\partial H}{\partial w^i} + \frac{X_i}{\beta}\frac{\partial H}{\partial U} \tag{50}$$

We then recover the second proposed equation for evolution of momentum:

$$-h\frac{dX_i}{dt} = k\{X_i, H\} = -\frac{k}{\beta}\frac{\partial H}{\partial w^i} + \frac{kX_i}{\beta}\frac{\partial H}{\partial U} \tag{51}$$

To reconcile the evolution of the internal energy with the third proposed equation, we recall the following thermodynamic identity:

$$\left(\frac{\partial H}{\partial \beta}\right)_{\mathbf{Y}} = \left(\frac{\partial H}{\partial \beta}\right)_{\mathbf{X}} + \sum_j \left(\frac{\partial H}{\partial X_j}\right)_{\beta, X_{i\neq j}} \left(\frac{\partial X_j}{\partial \beta}\right)_{\mathbf{Y}} \tag{52}$$

The subscripts indicate which parameters are held constant. This notation is common in thermodynamics and was previously omitted, as we specified at each step the symplectic form under consideration. Next, we find the partial derivative of the momentum with respect to the parameter β. From the coupling relation:

$$X_i = \frac{Y_i}{\beta} \Rightarrow \frac{\partial X_i}{\partial \beta} = -\frac{1}{\beta^2}Y_i = -\frac{1}{\beta}X_i \tag{53}$$

We can then obtain the evolution of the internal energy:

$$-h\frac{dU}{dt} = k\frac{\partial H}{\partial \beta} - \frac{1}{\beta}\sum_j \frac{\partial H}{\partial X_j}X_j \tag{54}$$

For the weight–momentum brackets, we note that:

$$\frac{\partial H}{\partial Y_i} = \frac{\partial H}{\partial X_i}\frac{\partial X_i}{\partial Y_i} = \frac{1}{\beta}\frac{\partial H}{\partial X_i} \tag{55}$$

Thus:

$$-h\frac{dw^i}{dt} = \frac{k}{\beta}\frac{\partial H}{\partial X_i} \tag{56}$$

Finally, the equation describing changes of H in the direction $\partial/\partial U$ remains unchanged, since the variable U is not coupled in the same manner as β.

References

[1] Herbert B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, 2 edition, 1985.

[2] Ana Cannas da Silva. *Lectures on Symplectic Geometry*, volume 1764 of *Lecture Notes in Mathematics*. Springer, 2001.

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

[4] Christopher M. Bishop and Hannah Bishop. *Deep Learning: Foundations and Concepts*. Springer International Publishing, Cham, 2024.

[5] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, 1992.

[6] John C. Baez. What is entropy?, 2024. arXiv preprint.

[7] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2 edition, 2013.

[8] William Rowan Hamilton. On a general method in dynamics. *Philosophical Transactions of the Royal Society of London*, 124:247–308, 1834.

[9] Robert G. Brown. *Exponential Smoothing: Forecasting and Control*. Prentice-Hall, 1956.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. arXiv preprint.

[11] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010*, pages 177–186, 2010.

[12] Pranav Vaidhyanathan, Lucas Schorling, Natalia Ares, and Michael A. Osborne. A physics-inspired optimizer: Velocity regularized adam, 2025. arXiv preprint.

[13] Guilherme França, Jeremias Sulam, Daniel P. Robinson, and Renato Vidal. Conformal symplectic and relativistic optimization, 2019. arXiv preprint.

[14] Dmitry Guskov and Vitaly Vanchurin. Covariant gradient descent, 2025. arXiv preprint.

[15] Geoffrey Hinton. Neural networks for machine learning, lecture 6e. Coursera (online course), 2012.