Theoretical Computer Science 633 (2016) 83–99

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

The principle of least cognitive action

Alessandro Betti [a], Marco Gori [b,*]

[a] Department of Physics, University of Pisa, Italy
[b] Department of Information Engineering and Mathematics, University of Siena, Italy

ARTICLE INFO

Article history:
Received 3 January 2015
Received in revised form 17 June 2015
Accepted 19 June 2015
Available online 2 July 2015

Keywords:
BG-brackets
Lifelong learning
Natural learning theory
On-line learning
Least cognitive action

ABSTRACT

By and large, the interpretation of learning as a computational process taking place in both humans and machines is primarily provided in the framework of statistics. In this paper, we propose a radically different perspective in which the emergence of learning is regarded as the outcome of laws of nature that govern the interactions of intelligent agents with their own environment. We introduce a *natural learning theory* based on the principle of *least cognitive action*, which is inspired to the related mechanical principle, and to the Hamiltonian framework for modeling the motion of particles. The introduction of the kinetic and of the potential energy leads to a surprisingly natural interpretation of learning as a dissipative process. The kinetic energy reflects the temporal variation of the synaptic connections, while the potential energy is a penalty that describes the degree of satisfaction of the environmental constraints. The theory gives a picture of learning in terms of the energy balancing mechanisms, where the novel notions of boundary and bartering energies are introduced. Finally, as an example of application of the theory, we show that the supervised machine learning scheme can be framed in the proposed theory and, in particular, we show that the Euler–Lagrange differential equations of learning collapse to the classic gradient algorithm on the supervised pairs.

1. Introduction

Machine Learning is at a lively crossroad of disciplines, where the exploration of neuroscience and cognitive psychology meets computational models mostly based on statistical foundations. For these models to be realistic, one is typically concerned with the acquisition of learning skills on a sample of training data, that is sufficiently large to be consistent with a statistically relevant test set. However, in most challenging and interesting learning tasks taking place in humans, the underlying computational processes do not seem to offer such a neat identification of the training set. As time goes by, humans react surprisingly well to new stimuli, while keeping past acquired skills, which seems to be hard to reach with nowadays intelligent agents. This suggests us to look for alternative foundations of learning, which are not necessarily based on statistical models of the whole agent life.

In this paper, we investigate the emergence of learning as the outcome of laws of nature that govern the interactions of intelligent agents with their own environment, regardless of their nature. The underlying principle is that the acquisition of cognitive skills by learning obeys information-based laws on these interactions, which hold regardless of biology. In this new perspective, in particular, we introduce a *natural learning theory* aimed at discovering the fundamental temporally-embedded

* Corresponding author.
E-mail address: marco@dii.unisi.it (M. Gori).

http://dx.doi.org/10.1016/j.tcs.2015.06.042
0304-3975/

mechanisms of environmental interaction. While the role of time has been quite relevant in machine learning (see e.g. the Hidden Markov Models), in some challenging problems, like computer vision, the temporal dimension seems to play quite a minor role. What could human vision had been in a world of visual information with shuffled frames? Any cognitive process aimed at extracting symbolic information from images that are not frames of a temporally coherent visual stream would have been extremely harder than in our visual experience [1]. More than looking for smart algorithmic solutions to deal with temporal coherence, in this paper we rely on the idea of regarding learning as a process which is deeply interwound with time, just like in most laws of nature.

In order to derive the laws of learning, we introduce the *principle of least cognitive action*, which is inspired to the related mechanical principle, and to the Hamiltonian framework for modeling the motion of particles. Unlike mechanics, however, the cognitive action that we define is in fact the objective to be minimized, more than a functional for which to discover a stationary point. In our learning framework, this duality is based on a proper introduction of the "kinetic" and of the "potential energy," that leads to a surprisingly natural interpretation of learning as a dissipative process. The kinetic energy reflects the temporal variation of the synaptic connections, while the potential energy is a penalty that describes the degree of satisfaction of the environmental constraints. The theory gives a picture of learning in terms of the energy balancing mechanisms, where the novel notions of *boundary and bartering energies* are introduced. These new energies arise mostly to model complex environmental interactions of the agent, that are nicely captured by means of time-variant high-order differential operators. When pairing them with the novel notion of BG-bracket, we show how intelligent processes can be understood in terms of energy balance; learning turns out be a dissipative process which reduces the potential energy by moving the connection weights, thus producing a kinetic energy. However, the energy balancing does involve also the boundary and bartering energies. The former arises because of the energy exchange at the beginning and at the end of the temporal horizon, while the last one is deeply connected with the energy exchange during the agent's life. It is worth mentioning that this energy balance, which has a nice qualitative interpretation, is in fact the formal outcome of the main theoretical results given in the paper. In particular, the proposed framework incorporates classic mechanics as a special case, while it opens the doors to in-depth analyses in any context in which there is an explicit temporal dependence in the Lagrangian of the system.

While the paper focuses on laws of learning that are independent of the nature of the intelligent agent, it is shown that these laws also offer a general framework to derive classic on-line machine learning algorithms. In particular, we show that the supervised machine learning scheme can be framed in the proposed theory, and that the Euler–Lagrange differential equations of learning derived in the paper do collapse to the classic gradient algorithm. Interestingly, the theory prescribes the time-variant learning rate, which arises from the given variational formulation.

The paper is organized as follows. In the next section, we introduce natural principles of learning and, in particular, the notion of cognitive action. Its minimization, which leads to the laws of learning, is described in Section 3. In Section 4 we propose a dissipative Hamiltonian framework of learning, which, in Section 5, is completed by the analysis of the BG-brackets. Section 6 relies on the previous results to provide an interpretation of learning dynamics by means of energy balance. In Section 7, we show how the laws of learning can be converted to the gradient algorithm. Finally, some conclusions are drawn in Section 8.

2. Natural principles of learning

The notion of time is ubiquitous in laws of nature. Surprisingly enough, most studies on machine learning have relegated time to the related notion of iteration step.[1] From one side the connection is sound, since it involves the computational effort, that is also observed in nature. From the other side, while time has a truly continuous nature, the time discretization typical in fields like computer vision seems to give up to the challenge of constructing a truly theory of learning. This paper presents an approach to learning that incorporates time in its truly continuous structure, so as the evolution of the weights of the neural synapses follows equations that resemble laws of physics. We consider environmental interactions that are modeled under the recently proposed framework of *learning from constraints* [2]. In particular, we focus attention on the case in which each constraint originates a loss function, that is expected to take on small values in case of soft-satisfaction. Interestingly, thanks to the adoption of the t-norm theory [3], loss functions can be constructed that can also represent the satisfaction of logic constraints.

A lot of emphasis in machine learning has been on *supervised learning* where we are given the collection $\mathcal{L} = \{(t_\kappa, u(t_\kappa)), s_\kappa\}_{\kappa\in\mathbb{N}}$ of supervised pairs $(u(t_\kappa), s_\kappa)$ over the temporal sequence $\{t_\kappa\}_{\kappa\in\mathbb{N}}$. At a given t, the pairs $(u(t_\kappa), s_k)$ that have become available are denoted by $\triangleright\mathcal{L}_t := \{\kappa \in \mathbb{N} : \ t_\kappa \leq t\}$. We assume that those data are learned by a *feedforward neural network* defined by the function

$$f(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^D,$$

so as the input $u(t)$ is mapped to $y(t) = f(w(t), u(t))$. Classic supervised learning can be naturally modeled by introducing a loss function $L(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ along with the given training set $\mathcal{L}$. Examples of different loss functions can be found

[1] This is often named "epoch."

Table 1
Links between the natural learning theory and classical mechanics.

Natural Learning Theory $\leftsquigarrow$ Mechanics	Remarks
$w_i \leftsquigarrow q_i$	Weights are interpreted as generalized coordinates.
$\dot{w}_i \leftsquigarrow \dot{q}_i$	Weights variations are interpreted as generalized velocities.
$\upsilon_i \leftsquigarrow p_i$	The conjugate momentum to the weights is defined by using the machinery of Legendre transforms.
$A[w] \leftsquigarrow S[q]$	The cognitive action is the dual of the action in mechanics.
$F(t, w, \dot{w}) \leftsquigarrow L(t, q, \dot{q})$	The Lagrangian F tis associated with the classic Lagrangian L in mechanics.
$H(t, w, \upsilon) \leftsquigarrow H(t, q, p)$	When using w and υ, we can define the Hamiltonian, just like in mechanics.

in [4] (Ch. 3). A classic case is the one of quadratic loss, that is $L(y, s) = \frac{1}{2}(y - s)^2$. Because of the variational formulation of learning that is given in this paper, it is convenient to associate L with the overall distributional loss

$$V(t, w) = \frac{1}{2}\psi(t) \sum_{\kappa \in \triangleright \mathcal{L}_t} (f(w(t), u(t)) - s_\kappa)^2 \cdot \delta(t - t_\kappa).$$

Here, $\psi(\cdot)$ is referred to as the *dissipation function*, and it is assumed to be a positive and increasing monotonic function on its domain. A fundamental principle behind the emergence of learning, which will be fully gained in the remainder of the paper, is that it does require energy dissipation. In the Hamiltonian framework, the introduction of dissipation processes can be done in different ways. The idea of factorizing the potential, as well as the whole Lagrangian, by a temporally-growing exponential term is related to studies on dissipative Hamiltonian systems [5]. We regard $w \in \mathbb{R}^n$ as the *Lagrangian coordinates* of a virtual mechanical system. Throughout this paper, $V(\cdot, \cdot)$ is referred to as the *potential energy* of the system defined by Lagrangian coordinates w. In machine learning, we look for trajectories $w(t)$ that possibly lead to configurations with small potential energy. Following the duality with mechanics, we also introduce the notion of kinetic energy. For reasons that will become clear in the remainder of the paper, we generalize the notion of velocity, $T_i w_i$, where, for each fixed i, T_i is a time dependent linear differential operator of the form T_i

$$T_i(t) = \sum_{j=0}^{\ell} \alpha_{i,j}(t) \frac{d^j}{dt^j}. \tag{1}$$

Let $m_i > 0$ be the *mass* of each particle (dual of connection weight) defined by the *position* $w_i(t)$ and *velocity* $\dot{w}_i$. Then, we define the *kinetic energy* as

$$K(t, Tw) = \frac{1}{2}\psi \sum_{i=1}^{n} m_i (T_i w_i)^2. \tag{2}$$

Clearly, if $T_i = T = d/dt$ and $\psi(t) \equiv 1$ then we have $K = \frac{1}{2}\sum_{i=1}^{n} m_i \dot{w}_i^2$, which returns the classic case of analytic mechanics. We notice on passing that one could always adsorb the dissipative function factor by introducing $T_i^{\psi} := \sqrt{\psi} T_i$, so as $K(t, Tw) = \frac{1}{2}\sum_{i=1}^{n} m_i (T_i^{\psi} w_i)^2$. Now, let us consider a Lagrangian simply composed of weighed the sum of the potential $V(\cdot, \cdot)$ and the kinetic energy

$$F(t, w, Tw) = \sum_{i=1}^{n} F(t, w_i, T_i w_i) = \sum_{i=1}^{n} K(t, T_i w_i) + \gamma V(t, w). \tag{3}$$

Here we mostly assume $\gamma \in \mathbb{R}$, so as the cognitive action gets a clear meaning in terms of regularization theory. However, it is important to notice that if $\gamma = -1$ the cognitive action strictly relates to classic action in mechanics. The consequences of different choices of γ will become more clear in the reminder of the paper. The classic problem of supervised learning is that of discovering $w^o = \arg\min A[w]$, where

$$A[w] = \int_{t_0}^{t_1} F(t, w, Tw)\, dt \tag{4}$$

is the *cognitive action* of the system. While there is an intriguing analogy with analytic mechanics (see Table 1), we emphasize that, while in mechanics we are only interested in stationary points of the mechanical action, in learning, we would also like to determine the minimum of cognitive action defined by equation (4). Interestingly, as we will see later, the strict connection with analytic mechanics also makes sense when considering strong dissipation (see also [6]).

Finally, a comment on the minimization of the cognitive action (4) is in order. In general, the problem makes sense whenever the Lagrangian is given all over $[t_0, t_1]$. The very nature of the problem results in the explicit time dependence of $F(\cdot, \cdot, \cdot)$, which, in turn, depends on the information coming from the interactions of the agent with the environment. As a

consequence, if there is no restriction on the nature of this interaction then the optimization problem does require all the information coming in the $[t_0, t_1]$ interval. However, the environmental interactions that are of interest in this paper are those that are typically associated with learning processes, where after the agent has inspected a certain amount of information, it makes sense to involve it in predictions on the future. Hence, whenever we deal with a truly *learning environment*, in which the agent is expected to capture regularities in the inspected data, as it will be shown, the minimization problem becomes well-posed and nicely fits in the context of on-line learning.

3. Euler–Lagrange laws of learning

Let us consider the minimization of functional $A[\cdot]$ defined by equation (4). Throughout the paper, we assume that the linear operators T admits the adjoint $T^\star$. As stated in the following proposition, the coefficients of $T^\star$ come directly from T's.

Proposition 3.1. *Given the linear differential operator T of degree ℓ, defined by (1), its adjoint operator is also a linear differential operator of the same degree with coefficients*

$$\beta_\kappa(t) = \sum_{j\geq\kappa}(-1)^j \binom{j}{\kappa} \alpha_j^{(j-\kappa)}(t) \tag{5}$$

Proof. If we use Leibniz formula, we have

$$\frac{d^j}{dt^j}(\alpha_j(t)\cdot\chi(t)) = \sum_{\kappa=0}^{j}\binom{j}{\kappa}\alpha_j^{(j-\kappa)}(t)\frac{d^\kappa\chi}{dt^\kappa}(t).$$

From the definition of adjoint operator, we can easily see that, if we pose $D^j := d^j/dt^j$, then its adjoint is $(D^j)^\star = (-1)^j D^j$. Then we get

$$\begin{aligned} T^\star(t) &= \sum_{j=0}^{\ell}(-1)^j \sum_{\kappa=0}^{j}\binom{j}{\kappa}\alpha_j^{(j-\kappa)}(t)\frac{d^\kappa}{dt^\kappa} \\ &= \sum_{\kappa=0}^{\ell}\sum_{j\geq\kappa}(-1)^j\binom{j}{\kappa}\alpha_j^{(j-\kappa)}(t)\frac{d^\kappa}{dt^\kappa} \equiv \sum_{k=0}^{\ell}\beta_k(t)\frac{d^k}{dt^k} \end{aligned}$$

from which the thesis follows. □

Let us calculate the variation of the functional (4), in the case $n = 1$, over the interval $[t_0, t_1]$ by assuming that[2] the stationary point $w(\cdot)$ has end-points $(t_0, w^{(\kappa)}(t_0))$ and $(t_1, w^{(\kappa)}(t_1))$ with $\kappa = 0, \ldots, \ell-1$. Let us consider $\check{w}(t) = w(t) + \epsilon\xi(t)$, where $\xi(\cdot)$ is a variation and $\epsilon \in \mathbb{R}$. Clearly, the conditions on the boundaries on $w^{(\kappa)}$ yield corresponding conditions on the variation

$$\xi^{(\kappa)}(t_0) = \xi^{(\kappa)}(t_1) = 0, \qquad \kappa = 0, \ldots, \ell-1. \tag{6}$$

Then we have

$$\begin{aligned} \delta A &= A[\check{w}] - A[w] \\ &= \int_{t_0}^{t_1}[F(t, w+\epsilon\xi, Tw+\epsilon T\xi) - F(t, w, Tw)]\,dt \\ &= \epsilon\cdot\int_{t_0}^{t_1}[F_w\cdot\xi + F_{Tw}\cdot T\xi]\,dt + \mathcal{O}(\epsilon^2). \end{aligned}$$

Now, because of conditions (6), the second term can be expressed by

$$\int_{t_0}^{t_1} F_{Tw}\cdot T\xi\,dt = \int_{t_0}^{t_1}(T^\star F_{Tw})\cdot\xi\,dt.$$

[2] Here, we use the κ-order derivative notation $D^\kappa w := w^{(\kappa)}$ and, for $\kappa = 0$, we assume $w^0 := w$.

Hence, we get

$$\delta A = \epsilon \cdot \int_{t_0}^{t_1} [F_w + T^\star F_{Tw}] \cdot \xi \, dt.$$

When imposing the stationary condition $\delta A = 0$, because of the fundamental lemma of variational calculus we get

$$F_w + T^\star F_{Tw} = 0. \tag{7}$$

Now we can easily see that the above result can be generalized to the set of variables w_i, $i = 1, \ldots, n$. Let us consider the general case in which we use an operator T_i for each variable w_i. In this case the variation turns out to be

$$\delta A = \epsilon \cdot \int_{t_1}^{t_2} \sum_{i=1}^{n} (F_{w_i} + T_i^\star F_{T_i w_i}) \cdot \xi_i \, dt.$$

Finally, this analysis leads to a necessary condition for the stationary of functional A, that is given in the following theorem.

Theorem 3.1. *Let us assume that we are given the values of $w^{(\kappa)}(t_0)$ and $w^{(\kappa)}(t_1)$ with $\kappa = 0, \ldots, \ell - 1$. Then functional (4) admits a stationary point w_i, $i = 1, \ldots, n$, provided that*

$$F_{w_i} + T_i^\star F_{T_i w_i} = 0 \qquad i = 1, \ldots, n. \tag{8}$$

Remark 3.1. The proof of the theorem follows the classic principles of variational calculus to determine stationary points (see e.g. [7,8]). The only specific issue that arises in the given result concerns the presence of operators T_i instead of the single derivatives, which plays an important role in the reminder of the paper.

Remark 3.2. In the formulation of learning, one cannot typically rely on the boundary conditions that are assumed in the theorem, whereas it makes sense to make assumptions on the initial conditions $w^{(\kappa)}(t_0)$, $\kappa = 0, \ldots, 2\ell - 1$ (Cauchy conditions). Interestingly, in case in which the Euler–Lagrange equations (8) are asymptotically stable, the effect of the initial condition vanishes as $t \to \infty$. In this paper we are mostly interested in exploring the environmental interactions of the agent under asymptotic stability.

Remark 3.3. Equations (8) have been derived under the assumption that we are given the end-points $(t_0, w^\kappa(t_0))$ and $(t_1, w^\kappa(t_1))$ with $\kappa = 0, \ldots, \ell - 1$. While they represent a necessary condition for stationarity, their integration requires the knowledge of the boundary condition, which leads to a batch-mode formulation of learning. This is related to the comments at the end of Section 2 concerning the on-line formulation of learning. However, if the learning environment is *periodic* with period τ and n_τ pairs in each period, that is if $(u(t_i), s_i) = (u(t_i + \tau), s_{i+n_\tau})$, there is experimental evidence to claim that the causality issue raised at the end of Section 2 disappears [9]. Basically, the on-line learning on periodic learning environment returns a solution that very well approximates the one corresponding to the batch-mode. A more interesting case, that is worth investigating, is the one of almost periodic [10] environments in which, roughly speaking, τ can depend on t. The restriction to *truly learning online environments*, in which the regularities can be captured because of underlining statistical assumptions, might be better captured by boundary conditions different with respect to those expressed by equations (6). For example, in the simplest case of $T = D$, one can impose that the weights converge over large intervals, which corresponds with the transversality condition [8] on the right border $F_{T_{w_i}}(t, w_i, T w_i)|_{t=t_1} \simeq 0$. However, in this paper, we do not address this issue, whereas we explore the behavior of equations (8) by means of an analysis based on "energy balance".

Finally, it is also worth mentioning that the given Euler–Lagrange equations get to stationary points and, therefore, depending the learning task, the actual minimization of the cognitive action might not be achieved. This is somewhat related to the classic framework of supervised learning, where the discovering of optimal solutions also in finite-dimensional spaces may led to suboptimal solutions [11].

In the case of the cognitive action defined by the Lagrangian of equation (3) we have

$$T_i^\star(\psi T_i w_i) + \gamma V_{w_i} = 0 \qquad i = 1, \ldots, n. \tag{9}$$

REDUCTION TO MECHANICS

Let us assume that $\forall i = 1, \dots, n: \quad T_i = T = D$ be, and let $\gamma = -1$ be. Then we consider the Lagrangian

$$F(t, w, Tw) = \frac{1}{2}\psi \sum_{i=1}^{n} m_i (Dw_i)^2 - \psi V(w)$$

where $\psi = e^{\theta t}$. Then the Euler–Lagrange equations (see Theorem 3.1) become

$$D^2 w_i + \theta D w_i + V_{w_i} = 0 \qquad i = 1, \dots, n, \tag{10}$$

which corresponds with damping oscillators in classic mechanics. Here we can promptly see the role of function $\psi(\cdot)$ in the birth of dissipation. Moreover, the role of $\gamma = -1$ becomes clear in order to attach the classic meaning of potential to function V. In addition, this is fully coherent with the definition of action, where the Lagrangian is $L = K - V$.

EVEN-ORDER γ SIGN FLIP

Let us consider the case $T = \alpha_o + \alpha_1 D + \alpha_2 D^2$ and $\psi(t) = e^{\theta t}$.

From Proposition 3.1, we have $T^\star = \alpha_o - \alpha_1 D + \alpha_2 D^2$ and we can easily see that the EL-equations becomes

$$D^4 w_i + \beta_3 D^3 w_i + \beta_2 D^2 w_i + \beta_1 D w_i + \beta_o w_i + \frac{\gamma}{\alpha_2^2 m_i} V_{w_i} = 0 \tag{11}$$

where

$$\beta_o := \frac{\alpha_o \alpha_2 \theta^2 - \alpha_o \alpha_1 \theta + \alpha_o^2}{\alpha_2^2}$$

$$\beta_1 := \frac{\alpha_1 \alpha_2 \theta^2 + (2\alpha_o \alpha_2 - \alpha_1^2)\theta}{\alpha_2^2}$$

$$\beta_2 := \frac{\alpha_2^2 \theta^2 + \alpha_1 \alpha_2 \theta + 2\alpha_o \alpha_2 - \alpha_1^2}{\alpha_2^2}$$

$$\beta_3 := 2\theta.$$

The comparison with equation (10) reveals and interesting difference which involves the role of the sign of γ in stability. While in mechanics we have $\gamma = -1$, when involving the second-order differential operator of this example, in order to get stability, from Routh–Hurwitz stability criterion we immediately conclude that $\gamma > 0$ is a necessary condition. This can be straightforwardly extended to any even order of T. In this paper, this property is referred to as γ sign flip, and motivates the study of high-order operators.

4. Dissipative Hamiltonian formulation of learning

Now we use Legendre transform in order to obtain the Hamiltonian formulation of the problem. We introduce the notations $\dot{f}_i := T_i f_i$ and $\mathring{f}_i := T_i^\star f_i$, so as equations (8) can be re-written as

$$F_{w_i} + \mathring{F}_{\dot{w}_i} = 0 \quad i = 1, 2, \dots, n. \tag{12}$$

Now, let us introduce the *conjugate momentum* υ_i

$$\upsilon_i := \frac{\partial F}{\partial \dot{w}_i}. \tag{13}$$

Likewise, we define the Hamiltonian as

$$H(t, w, \upsilon) := \sum_{i=1}^{n} \upsilon_i \dot{w}_i - F(t, w, \dot{w}), \tag{14}$$

where $\dot{w}_i$ must be thought of as functions of w_i and υ_i.

Theorem 4.1. *The Euler–Lagrange equations (8) are equivalent to the* 2*n Hamilton equations*

$$\dot{w}_i = \frac{\partial H}{\partial \upsilon_i}, \tag{15}$$

$$\mathring{\upsilon}_i = \frac{\partial H}{\partial w_i}. \tag{16}$$

Moreover, we have

$$\frac{\partial H}{\partial t} = -\frac{\partial F}{\partial t}. \tag{17}$$

Proof. As usual, we introduce the canonical variables w_i and υ_i. Then the canonical equations arise from the EL-equations (8) and from definition (13), as follows:

$$\frac{\partial H}{\partial \upsilon_i} = \frac{\partial}{\partial \upsilon_i}\left(\sum_{j=1}^{n} \upsilon_j \dot{w}_j - F(t, w, \dot{w})\right) = \dot{w}_i + \upsilon_i \frac{\partial \dot{w}_i}{\partial \upsilon_i} - \frac{\partial F}{\partial w_i}\frac{\partial w_i}{\partial \upsilon_i} - \frac{\partial F}{\partial \dot{w}_i}\frac{\partial \dot{w}_i}{\partial \upsilon_i} = \dot{w}_i,$$

$$\frac{\partial H}{\partial w_i} = \frac{\partial}{\partial w_i}\left(\sum_{j=1}^{n} \upsilon_j \dot{w}_j - F(t, w, \dot{w})\right) = \dot{w}_i \frac{\partial \upsilon_i}{\partial w_i} + \upsilon_i \frac{\partial \dot{w}_i}{\partial w_i} - \frac{\partial F}{\partial w_i} - \frac{\partial F}{\partial \dot{w}_i}\frac{\partial \dot{w}_i}{\partial w_i}$$

$$= -\frac{\partial F}{\partial w_i} = T^{\star}\frac{\partial F}{\partial \dot{w}_i} = \mathring{\upsilon}_i.$$

Finally, equation (17) arises from (16) as follows

$$\frac{\partial H}{\partial t} = \frac{\partial}{\partial t}\left(\sum_{i=1}^{n} \upsilon_i \dot{w}_i - F(t, w, \dot{w})\right) = \sum_{i=1}^{n} \upsilon_i \frac{\partial \dot{w}_i}{\partial t} - \sum_{i=1}^{n} F_{\dot{w}_i}\frac{\partial \dot{w}_i}{\partial t} - \frac{\partial F}{\partial t}$$

$$= -\frac{\partial F}{\partial t}. \qquad \square$$

The integration of Hamilton equations (15) and (16) can be written formally as

$$\begin{aligned} w_i &= \overset{-1}{T}_i H_{\upsilon_i} \\ \upsilon_i &= \overset{-1}{T}{}^{\star}_i H_{w_i}, \end{aligned} \tag{18}$$

where $\overset{-1}{T}_i$ and $\overset{-1}{T}{}^{\star}_i$ denote the inverse operators of T_i and $T^{\star}_i$, respectively. Notice that whenever we use the inverse operator, we need to solve an ordinary differential equation of order ℓ and, therefore, we must always impose ℓ initial conditions to get a unique solution. While equations (15) and (16) can be used for determining the evolution of w, as it will be shown in Section 6, equation (15), can offer a picture of learning in terms of energy balance. The following definition plays a fundamental role in the analysis of energy balance. The definition involves the set of variables $\mathcal{N} := \{(w_i, \upsilon_i),\ i \in \mathbb{N}_n\}$ paired with the associated differential operator T.

Definition 4.1. Given the system $\Sigma \sim (\mathcal{N}, T)$, along with any pair of functions $\phi_1(t, w, \upsilon)$ and $\phi_2(t, w, \upsilon)$, with continuous derivatives up to the ℓ-th order, the term

$$\{\phi_1, \phi_2\}^T_D := \sum_{i=1}^{n}\left(\frac{\partial \phi_1}{\partial w_i} D \overset{-1}{T}\frac{\partial \phi_2}{\partial \upsilon_i} + \frac{\partial \phi_1}{\partial \upsilon_i} D \overset{-1}{T^{\star}}\frac{\partial \phi_2}{\partial w_i}\right) \tag{19}$$

is referred to as the bracket of ϕ_1 and ϕ_2, with respect to operators D and T.

For the sake of simplicity, in the reminder of the paper, we will omit the subscript and superscript, so as we will simply use the nation $\{\phi_1, \phi_2\}$. We also notice that this can be regarded as a formal definition, unless we know (as it will be in what follows) the initial conditions for

$$\overset{-1}{T}\frac{\partial \phi_2}{\partial \upsilon_i} \quad \text{and} \quad \overset{-1}{T^{\star}}\frac{\partial \phi_2}{\partial w_i}.$$

We can easily see that general $\{\phi_1, \phi_2\} \neq \{\phi_2, \phi_1\}$ and that $\{\phi_1, \phi_2\}$ is linear only in the first argument.

Theorem 4.2. *Given the system defined by $w : [t_0, t_1] \to \mathbb{R}^n$ and any $\phi(t, w, \upsilon)$ with continuous derivates up to the ℓ-th order, we have*

$$\frac{d\phi}{dt} = \frac{\partial \phi}{\partial t} + \{\phi, H\}. \tag{20}$$

Proof. We have

$$\frac{d\phi}{dt} = \frac{\partial \phi}{\partial t} + \sum_{i=1}^{n} \left(\phi_{w_i} D w_i + \phi_{\upsilon_i} D \upsilon_i \cdot \right) \tag{21}$$

If we plug w_i and υ_i given by equation (18) into equation (21) we get

$$\frac{d\phi}{dt} = \frac{\partial \phi}{\partial t} + \sum_{i=1}^{n} \left(\phi_{w_i} D \overset{-1}{T} H_{\upsilon_i} + \phi_{\upsilon_i} D \overset{-1}{T^\star} H_{w_i} \right) = \frac{\partial \phi}{\partial t} + \{\phi, H\}. \qquad \square$$

5. BG-brackets

In this section we describe some properties of the brackets $\{\cdot,\cdot\}$ introduced by Definition 4.1. Their analysis leads to an in-depth understanding of energy exchange processes. We start noticing that if $\forall i = 1, \ldots, n: \ T_i = D$ then $\{\cdot,\cdot\}$ reduce to classic Poisson's brackets. This is formally stated by the following proposition

Proposition 5.1. *The brackets $\{\cdot,\cdot\}$ are a generalization of Poisson brackets to the case in which $\forall i = 1, \ldots, n: \ T_i = D$, that is*

$$\{\phi_1, \phi_2\} = \{\phi_1, \phi_2\}_{P.B.}.$$

Proof. From Proposition 5, under the assumption that $\forall i = 1, \ldots, n: \ T_i = D$, we derive that $T_i^\star = -D$. Hence, we have

$$\begin{aligned} \{\phi_1, \phi_2\} &= \sum_{i=1}^{n} \left(\frac{\partial \phi_1}{\partial w_i} D \overset{-1}{D} \frac{\partial \phi_2}{\partial \upsilon_i} - \frac{\partial \phi_1}{\partial \upsilon_i} D \overset{-1}{D} \frac{\partial \phi_2}{\partial w_i} \right) \\ &= \sum_{i=1}^{n} \left(\frac{\partial \phi_1}{\partial w_i} \frac{\partial \phi_2}{\partial \upsilon_i} - \frac{\partial \phi_1}{\partial \upsilon_i} \frac{\partial \phi_2}{\partial w_i} \right) \\ &= \{\phi_1, \phi_2\}_{P.B.}. \qquad \square \end{aligned}$$

REST OF THE ADJOINT

Now, we focus on the case[3] $T \neq D$. Interestingly, we will show that the enrichment of the differential operator T with respect to D has important consequences on the energy balance equation (20). While from Proposition 5.1 we have $\{H, H\} = 0$, it will be shown that $T \neq D$, in general leads to $\{H, H\} \neq 0$. Let

$$\langle w, \upsilon \rangle := \int_{t_0}^{t} w(\tau) \upsilon(\tau) \, d\tau \tag{22}$$

be. We introduce a couple of concepts that turn out to be useful in the following. First, we notice that any differential operator T can be paired with a real number, which comes out when considering the behavior on the borders of $\langle T w, \upsilon \rangle$.

Definition 5.1. Let T be a differential operator and let $\langle \cdot, \cdot \rangle$ be the inner product given by (22). Then for any given function pair of functions $f, g \in W^{\ell,2}$

$$\Lambda(T \,|\, f, g) := \langle Tf, g \rangle - \langle f, T^\star g \rangle, \tag{23}$$

is referred to as the *rest of the adjoint* of T with respect to f and g.

Clearly, $\Lambda(\cdot \,|\, \cdot, \cdot)$ is linear in each of its arguments and it is distributive in both the operator and the two function slots. Moreover, we have that $\Lambda(c \cdot I \,|\, f, g) \equiv 0$ for any constant c. Notice that $\Lambda(T \,|\, f, g) := \langle Tf, g \rangle = 0$ whenever we are given the same boundary conditions on f and g.

Proposition 5.2. *Let $T^m = \sum_{i=0}^{m} \alpha_o D^i$ be, where α_i are real constant. Then the rest of the adjoint $\Lambda(T \,|\, w, v)$ w.r.t. w and υ can be computed by the recurrent equations*

[3] For the sake of simplicity, in the remainder of the paper, we assume $\forall i = 1, \ldots, n: \ T_i = T$.

$$
\begin{aligned}
&i. \quad \Lambda(T^m|\, w, v) = \Lambda(T^{m-1}|\, w, v) + \alpha_m \Lambda(D^m|\, w, v),\\
&ii. \quad \Lambda(D^m|\, w, v) = \left[v D^{m-1} w\right] - \Lambda(D^{m-1}|\, w, Dv),
\end{aligned} \tag{24}
$$

where $[u] := u(t_1) - u(t_0)$ *is the variation on the borders of* u.

Proof.

i. We have

$$
\begin{aligned}
\Lambda(T^m|\, w, v) &= \langle T^{m-1} w + \alpha_m D^m w, v\rangle - \langle w, (T^{m-1})^\star v + \alpha_m (D^m)^\star v\rangle\\
&= \Lambda(T^{m-1}|\, w, v) + \alpha_m \left(\langle D^m w, v\rangle - \langle w, (D^m)^\star v\rangle\right)\\
&= \Lambda(T^{m-1}|\, w, v) + \alpha_m \Lambda(D^m|\, w, v).
\end{aligned}
$$

ii. In order to get the recurrent equation *ii.*, we start using integration by parts as follows:

$$
\begin{aligned}
\Lambda(D^m w|\, v) &= \langle D^m w, v\rangle - \langle w, (D^m)^\star v\rangle\\
&= \langle D^m w, v\rangle - (-1)^m \langle w, (D^m) v\rangle\\
&= [v D^{m-1} w] - \langle Dv, D^{m-1} w\rangle + (-1)^{m-1} \left([w D^{m-1} v] - \langle Dw, D^{m-1} v\rangle\right)\\
&= [v D^{m-1} w] + [w (D^{m-1})^\star v] - \left(\langle Dv, D^{m-1} w\rangle + \langle Dw, (D^{m-1})^\star v\rangle\right).
\end{aligned}
$$

Now we have

$$
\Lambda(D^{m-1}|\, w, Dv) = \langle D^{m-1} w, Dv\rangle - \langle w, (D^{m-1})^\star D v\rangle,
$$

and, therefore, we get

$$
\begin{aligned}
\Lambda(D^m|\, w, v) &= [v D^{m-1} w] + [w (D^{m-1})^\star v]\\
&\quad - \left(\langle w, (D^{m-1})^\star D v\rangle + \Lambda(D^{m-1}|\, w, Dv) + \langle Dw, (D^{m-1})^\star v\rangle\right)\\
&= [v D^{m-1} w] + [w (D^{m-1})^\star v] - [w (D^{m-1})^\star v] - \Lambda(D^{m-1}|\, w, Dv)\\
&= [v D^{m-1} w] - \Lambda(D^{m-1}|\, w, Dv). \qquad \square
\end{aligned}
$$

Example 5.1. Let $T = \alpha_0 + \alpha_1 D + \alpha_2 D^2$. To calculate the rest of the adjoint we use Proposition 5.2. From equation (24)-*ii* we can calculate $\Lambda(D^m|\, w, v)$ for $m = 1, 2, 3$. First, notice that $\Lambda(D^0|\, w, v) = 0$. Then, for D we have

$$
\Lambda(D|\, w, v) = [vw] - \Lambda(D^0|\, w, D^0 v) = [vw].
$$

Now we can calculate $\Lambda(D^2|\, w, v)$:

$$
\begin{aligned}
\Lambda(D^2|\, w, v) &= [vDw] - \Lambda(D|\, w, v)\\
&= [vDw] - \Lambda(D|\, w, Dv)\\
&= [vDw] - [wDv] = [vDw - wDv].
\end{aligned}
$$

Finally, from equation (24)-*ii*, we get

$$
\Lambda(T|\, w, v) = [\alpha_1 wv + \alpha_2 (vDw - wDv)]. \tag{25}
$$

The notion of rest of the adjoint plays an important role in the remainder of the paper, where we are interested in the following extended definition which involves learning system $\Sigma \sim (\mathcal{N}, T)$.

Definition 5.2. Given $\Sigma \sim (\mathcal{N}, T)$ we define

$$
M(\Sigma, [t_0, t_1]) := \sum_{i=1}^{n} \Lambda(T|\, w_i, Dv_i) + \Lambda(D|\, w_i, T^\star v_i). \tag{26}
$$

$M(\Sigma, [t_0, t_1])$ is referred to as the rest of T with respect to $\mathcal{N}$.

The following lemma offers an appropriate re-writing of $M(\Sigma, [t_0, t_1])$ to gain interesting conclusions.

Lemma 5.1.

$$M(\Sigma, [t_0, t_1]) := \sum_{i=1}^{n} \langle (T + D)w_i, (T^\star + D)v_i \rangle. \tag{27}$$

Proof. The proof is a straightforward consequence of the definition. □

First, we notice that if $T = D$ then $T^\star + D = 0$ and, consequently $M(\Sigma, [t_0, t_1]) = 0$. Clearly, the lemma enlightens the way $M(\Sigma, [t_0, t_1])$ emerges from breaking the adjoint property $T^\star + D = 0$, that holds for $T = D$.

In the case of Example 5.1, using equation (25) we get

$$\begin{aligned} M(\Sigma, [t_0, t_1]) &= \sum_{i=1}^{n} \Lambda(T \mid w_i, Dv_i) + \Lambda(D \mid w_i, T^\star v_i) \\ &= \sum_{i=1}^{n} [\alpha_1 w_i Dv_i + \alpha_2 (Dv_i Dw_i - w_i D^2 v_i)] + [w_i(\alpha_0 - \alpha_1 Dv_i + \alpha_2 D^2 v_i)] \\ &= \sum_{i=1}^{n} [\alpha_0 w_i + \alpha_2 Dw_i Dv_i] \end{aligned}$$

This example suggests that $M(\Sigma, [t_0, t_1])$ depends on the value assumed by

$$B(t, w, Dw, Dv) = \sum_{i=1}^{n} \alpha_0 w_i + \alpha_2 Dw_i Dv_i$$

on the boundaries of $[t_0, t_1]$. We have

$$M(\Sigma, [t_0, t_1]) = B(t_1, w, Dw, Dv) - B(t_0, w, Dw, Dv).$$

Clearly, this holds for Σ with any differential operator T. As a consequence, whenever we know that Σ exhibits a periodic behavior on $[t_0, t_1]$ then

$$M(\Sigma, [t_0, t_1]) = 0.$$

Now we show a property of M which leads to a clear interpretation in terms of the energy exchanges of the agent with the environment.

Theorem 5.1. *Let T be a positive self-adjoint operator with constant coefficients and let us consider a system Σ such that $v_i = Tw_i$. Then $M \geq 0$.*

Proof. From the hypothesis $T = T^\star$. Then, from the definition of M we get

$$\begin{aligned} M(\Sigma, [t_0, t_1]) &= \sum_{i=1}^{n} \Lambda(T \mid w_i, Dv_i) + \Lambda(D \mid w_i, Tv_i) \\ &= \sum_{i=1}^{n} \langle Tw_i, Dv_i \rangle - \langle w_i, TDv_i \rangle + \langle Dw_i, Tv_i \rangle + \langle w_i, DTv_i \rangle \\ &= \sum_{i=1}^{n} \langle Tw_i, Dv_i \rangle + \langle Dw_i, Tv_i \rangle \\ &= \sum_{i=1}^{n} \langle (T + D)w_i, (T + D)v_i \rangle. \end{aligned}$$

Since $v_i = Tw_i$ and $T \geq 0$ we get

$$\begin{aligned} M(\Sigma, [t_0, t_1]) &= \sum_{i=1}^{n} \langle (T + D)w_i, (T + D)Tw_i \rangle \\ &= \sum_{i=1}^{n} \langle \underbrace{(T + D)w_i}_{u_i}, T \underbrace{(T + D)w_i}_{u_i} \rangle = \sum_{i=1}^{n} \langle u_i, Tu_i \rangle \geq 0. \end{aligned}$$

□

T–D Commutator

Another important asymmetry arises which plays an important role in case of time-dependent differential operators. The following definition turns out to be useful.

Definition 5.3. Given any two differential operators T_1 and T_2 we define

$$[T_1, T_2] := T_1 T_2 - T_2 T_1. \tag{28}$$

The differential operator $[T_1, T_2]$ is referred to as the *commutator* of T_1 and T_2.

Of course, we have $[T_1, T_2] + [T_2, T_1] = 0$. In particular, in the following, we are interested in the case in which $T_1 = T$ and $T_2 = D$. In that case

$$\begin{aligned} TDy - DTy &= \sum_{i=0}^{\ell} \alpha_i D^i Dy - D \sum_{i=1}^{\ell} \alpha_i D^i y \\ &= \sum_{i=0}^{\ell} \alpha_i D^{i+1} y - \sum_{i=0}^{\ell} D\alpha_i D^i y - \sum_{i=0}^{\ell} \alpha_i D^{i+1} y \\ &= -\sum_{i=0}^{\ell} D\alpha_i D^i y \equiv -\sum_{i=1}^{\ell} \beta_i D^i y. \end{aligned} \tag{29}$$

Clearly, $[T, D] = 0$ for constant α_i coefficients, whereas in general, $[T, D]$ is a differential operator defined by coefficients β_i. Like for the rest of the adjoint, it turns out to be useful to extend the notion of commutator to the case of a set conjugate variables $\mathcal{N}$ for differential operators T and D.

Definition 5.4. Let us consider the system $\Sigma = \{\mathcal{N}, T\}$ over $[t_0, t_1]$. Then

$$\Upsilon(\Sigma, [t_0, t_1]) := \sum_{i=1}^{n} \langle w_i, [T^\star, D] \upsilon_i \rangle \tag{30}$$

is referred to as the $T^\star - D$ commutator of $\mathcal{N}$.

Again, the T–D commutator expresses the degree of asymmetry between T and D. We have symmetry $[T, D] = TD - DT = 0$ whenever T is a time-invariant differential operator, whereas the asymmetry just arises because of the temporal evolution of coefficients α_i. In order to give an interpretation of $\Upsilon(\Sigma, [t_0, t_1])$, let us consider the case in which T is replaced with ϕT, where ϕ is a positive increasing monotonic function. This case has been already mentioned in Section 2 with $\phi(t) = \sqrt{\psi(t)}$. We discuss the basic idea in the simplest case in which $T = \phi D$.

Theorem 5.2. *Let us assume that $M(\Sigma, [t_0, t_1]) = 0$. If $T = \phi D$ then*

$$\Upsilon(\Sigma, [t_0, t_1]) = -\sum_{i=1}^{n} m_i \int_{t_0}^{t_1} (D w_i)^2 \phi D\phi \, dt. \tag{31}$$

Proof. First, we can easily see that

$$T^\star = -D\phi \cdot D^0 - \phi D,$$

where $D^0 = I$ is the identity operator. Now we can calculate

$$\begin{aligned} [T^\star, D] &= T^\star D - D T^\star = (-D\phi \cdot I - \phi D) D - D(-D\phi \cdot I - \phi D) \\ &= D^2 \phi \, I + D\phi \, D. \end{aligned}$$

Now, since $M(\Sigma, [t_0, t_1]) = 0$ and $\upsilon_i = m_i T w_i = m_i \phi D w_i$, we have

$$\begin{aligned} \Upsilon(\Sigma, [t_0, t_1]) &= \sum_{i=1}^{n} \int_{t_0}^{t_1} w_i (\upsilon_i D^2 \phi + D\phi \, D\upsilon_i) dt = \sum_{i=1}^{n} \int_{t_0}^{t_1} w_i (D(\upsilon_i D\phi)) dt \\ &= -\sum_{i=1}^{n} \int_{t_0}^{t_1} \upsilon_i D w_i D\phi \, dt = -\sum_{i=1}^{n} m_i \int_{t_0}^{t_1} D\phi (D w_i)^2 \phi \, dt. \end{aligned}$$

□

Let us consider the case in which $\psi(t) = \phi^2(t) = e^{\theta t}$. We get

$$-\Upsilon(\Sigma, [t_0, t_1]) = \sum_{i=1}^{n} m_i \int_{t_0}^{t_1} D\phi (Dw_i)^2 \phi \, dt = \frac{1}{2} \sum_{i=1}^{n} m_i \theta \int_{t_0}^{t_1} (Dw_i)^2 \psi \, dt > 0.$$

This offers a clear interpretation of $-\Upsilon(\Sigma, [t_0, t_1])$, which turns out to be the dissipated energy of Σ over $[t_0, t_1]$. However, notice that this interpretation is limited to the case $D\phi > 0$.

Example 5.2. Let $T = \sin(\omega t) \cdot D$ be. We have

$$T^\star = -\omega \cos(\omega t) \cdot I - \sin(\omega t)\ D$$

and

$$[T^\star, D] = -\omega^2 \sin(\omega t) \cdot I + \omega \cos(\omega t)\ D.$$

Under the assumption that $M(\Sigma, [t_0, t_1]) = 0$ and $\upsilon_i = m_i T w_i = m_i \sin \omega t \cdot D$ we have

$$\begin{aligned}\Upsilon(\Sigma, [t_0, t_1]) &= \sum_{i=1}^{n} \int_{t_0}^{t_1} w_i D(\omega \cos(\omega t) \upsilon_i) dt = -\sum_{i=1}^{n} \int_{t_0}^{t_1} \cos(\omega t) \upsilon_i D w_i dt \\ &= -\frac{1}{2} \sum_{i=1}^{n} m_i \int_{t_0}^{t_1} \sin 2\omega t (Dw_i)^2 dt.\end{aligned}$$

This example clearly shows that the sign of $\Upsilon(\Sigma, [t_0, t_1])$ can flip with functions $\psi(\cdot)$ that are not monotonic. Hence, $\Upsilon(\Sigma, [t_0, t_1])$ models interactions of the agent with the environment where the energy flows in both directions.

The $\{\cdot, \cdot\}$ brackets assume a special meaning when involving the Hamiltonian of a given system.[4]

Theorem 5.3. *Given the system Σ over $[t_0, t_1]$ then*

$$\{H, H\} = \frac{d}{dt}(M + \Upsilon). \tag{32}$$

Proof. Because of the Hamilton equations, we have

$$\begin{aligned}\{H, H\} &= \sum_{i=1}^{n} \left(H_{w_i} D \overset{-1}{T} H_{\upsilon_i} + H_{\upsilon_i} D \overset{-1}{T^\star} H_{w_i} \right) \\ &= \sum_{i=1}^{n} \left(T^\star \upsilon_i D \overset{-1}{T} T w_i + T w_i D \overset{-1}{T^\star} T^\star \upsilon_i \right) \\ &= \sum_{i=1}^{n} \left(T^\star \upsilon_i \cdot D w_i + T w_i \cdot D \upsilon_i \right).\end{aligned}$$

Now, if we integrate over $[t_0, t]$ and use the definition of rest of the adjoint and of commutator, we get

$$\begin{aligned}\int_{t_0}^{t} \{H, H\} d\tau &= \sum_{i=1}^{n} \int_{t_0}^{t} (T^\star \upsilon_i \cdot D w_i + T w_i \cdot D \upsilon_i) \, d\tau \\ &= \sum_{i=1}^{n} \langle T^\star \upsilon_i, D w_i \rangle + \langle T w_i, D \upsilon_i \rangle \\ &= \sum_{i=1}^{n} \Lambda(D | w_i, T^\star \upsilon_i) - \langle D T^\star \upsilon_i, w_i \rangle + \Lambda(T | w_i, D \upsilon_i) + \langle w_i, T^\star D \upsilon_i \rangle\end{aligned}$$

[4] Whenever there is no risk of ambiguity, in the following we drop the dependence on $\Sigma, [t_0, t_1]$ in M and Υ.

$$= \sum_{i=1}^{n} \Lambda(D|w_i, T^\star v_i) + \Lambda(T|w_i, Dv_i) + \sum_{i=1}^{n} \langle w_i, (T^\star D - DT^\star)v_i \rangle$$
$$= M + \Upsilon.$$

Finally, we get the thesis when deriving w.r.t. to t both sides. □

This result shows that the $\{\cdot,\cdot\}$ brackets assumes a special meaning in the case in which both the operands are the Hamiltonian function. In particular, in that case, there is a *Boundary Generation* expressed by $M(\Sigma, [t_0, t_1])$, due to the rest of the adjoint, and a *Bartering Generation* expressed by $\Upsilon(\Sigma, [t_0, t_1])$, which is due to the exchange of the operators $T^\star$ and D. For this reason, $\{\cdot,\cdot\}$ is referred to as the *BG-brackets*.

6. Energy balance in learning processes

In this section we discuss energy balancing connected with the general form of cognitive action whose Lagrangian is given by equation (3). While equations (15) and (16) can be used to determine the learning dynamics, equation (17) offers an interesting view of learning in terms of energy balancing. Notice that the equation involves the canonical variables w and v in H, as well as $\dot{w}$ in F. An alternative way of constructing energy balance only based on canonical variables is that of using expressing $\{H, H\}$ according to Theorem 5.3. We have

$$\frac{dH}{dt} = \frac{dK}{dt} - \gamma \frac{dV}{dt} = \frac{\partial H}{\partial t} + \{H, H\} = -\gamma \frac{\partial V}{\partial t} + \frac{\partial K}{\partial t} + \frac{d}{dt}(M + \Upsilon).$$

We also have

$$\frac{dV}{dt} = \frac{\partial V}{\partial t} + \{V, H\},$$
$$\frac{dK}{dt} = \frac{\partial K}{\partial t} + \{K, H\},$$

which suggests the introduction of functions $\mathcal{V}$ and $\mathcal{K}$ defined by

$$\frac{d\mathcal{V}}{dt} = \{V, H\} = \frac{dV}{dt} - \frac{\partial V}{\partial t}, \tag{33}$$
$$\frac{d\mathcal{K}}{dt} = \{K, H\} = \frac{dK}{dt} - \frac{\partial K}{\partial t}. \tag{34}$$

Finally, the analysis is summarized by the following theorem, which establishes the energy balance of any learning system. Unlike equation (17), the given energy balance is fully expressed in terms of canonical variables w and v.

Theorem 6.1. *The system (Σ, T) evolves according to the invariant condition*

$$\frac{d}{dt}(\mathcal{K} - \gamma\mathcal{V} - M - \Upsilon) = 0. \tag{35}$$

An important source of interaction is represented by the *environmental potential energy*

$$\mathcal{E} = \mathcal{E}(\Sigma, [t_0, t]) := \int_{t_0}^{t} \frac{\partial V}{\partial \tau} d\tau,$$

and by the bartering energy Υ, which appears whenever $T^\star$ does not commutate with D. While $\mathcal{K}$ reflects the velocity of the connections dynamics, the boundary energy M takes into account the exchange of energy at the beginning and at the current time of the learning horizon.

DAMPING OSCILLATORS

In order to provide an interpretation of analytic mechanics as a special case, let us consider $T = D$ and $\gamma = -1$. Basically, we are considering the Lagrangian which has originated the damping oscillator of equation (10) From the definition (30) of Υ we have $[T^\star, D] = [-D, D] = 0$ and, therefore, $\Upsilon = 0$. Likewise, from Lemma 5.1, since $D + T^\star = 0$, we conclude that $M = 0$. As a consequence, the invariant of Theorem 6.1, expressed in terms of canonical variables, yields

$$\frac{d}{dt}(\mathcal{K} + \mathcal{V}) = 0. \tag{36}$$

This is in fact the classic invariant condition of classic mechanics, which establishes that the sum of the kinetic and potential energies is invariant. Interestingly, this holds also in cases in which the kinetic energy K and the potential V are

dependent on time, since it is intimately connected with the nullification of $\{H, H\}$. In the classic case in which there is no dissipation, from the definition of $\mathcal{V}$ and $\mathcal{K}$, given in equations (33) and (34), this means that $\partial V/\partial t = 0$ and $\partial K/\partial t = 0$. As a consequence $d\mathcal{V}/dt = dV/dt$ and $d\mathcal{K}/dt = dK/dt$ which, in turn, yields

$$\frac{d}{dt}(K+V) = 0. \tag{37}$$

In case of dissipation, only the energy balance (36) holds, while the above balance clearly fails. In order to model the damping oscillations, following what has already been introduced in Section 3 we have $\upsilon_i = m_i \psi \dot{w}_i$, where $\psi(t) = e^{\theta t}$. Hence, we can express the potential and kinetic energy by

$$\begin{aligned} V(t,w) &= \psi \overline{V}(t,w), \\ K(t,\upsilon) &= \frac{1}{\psi}\sum_{i=1}^{n}\frac{\upsilon_i^2}{2m_i} = \frac{\overline{K}(\upsilon)}{\psi}. \end{aligned} \tag{38}$$

From the definition of $\mathcal{V}$ and $\mathcal{K}$ we promptly get

$$\begin{aligned} \frac{d\mathcal{V}}{dt} &= \psi \frac{d\overline{V}}{dt}, \\ \frac{d\mathcal{K}}{dt} &= \frac{1}{\psi}\frac{d\overline{K}}{dt}. \end{aligned}$$

Hence, the energy balance turns out to be

$$\frac{d}{dt}\overline{K}(\upsilon(t)) + \psi^2 \frac{d}{dt}\overline{V}(t, w(t)) = 0. \tag{39}$$

When $\psi(t) = e^{\theta t}$ we have $\psi(t) \geq 1$, which indicates that as the time increases, the kinetic energy drops more that in the case $\psi(t) = 1$ (no dissipation). A related energy balance can be obtained which only involves the weights. In this case let us define

$$\begin{aligned} K(t,\dot{w}) &= \frac{1}{2}\psi\sum_{i=1}^{n} m_i \dot{w}_i = \psi \overline{K}(\dot{w}), \\ V(t,w) &= \psi \overline{V}(t,w). \end{aligned}$$

It is worth mentioning that since equation (17) involves both the Hamiltonian H and the Lagrangian F, it is opportune to pose $\tilde{H}(t,w,\dot{w}) := H(t,w,\upsilon)$, where $\dot{w}$ and υ are connected by $\upsilon_i = m_i \psi \dot{w}_i$. Hence, we get[5]

$$\begin{aligned} DH = D(K+V) &= D(\psi\overline{K} + \psi\overline{V}) = \dot{\psi}\overline{K} + \psi\frac{d\overline{K}}{dt} + \dot{\psi}\overline{V} + \psi\frac{d\overline{V}}{dt} \\ &= -\frac{\partial F}{\partial t} = -\frac{\partial}{\partial t}(\psi\overline{K} - \psi\overline{V}) = -\dot{\psi}\overline{K} + \dot{\psi}\overline{V} + \frac{\partial\overline{V}}{\partial t}. \end{aligned}$$

Since $\psi(t) = e^{\theta t}$ we get

$$\frac{d\overline{K}}{dt} + 2\theta\overline{K} = -\frac{d\overline{V}}{dt} + \frac{\partial\overline{V}}{\partial t}. \tag{40}$$

If we integrate over the interval $[t_0, t_1]$ we get

$$\overline{K}(t_0) + \overline{V}(t_0) + \int_{t_0}^{t_1}\frac{\partial\overline{V}}{\partial t}dt = \overline{K}(t_1) + \overline{V}(t_1) + 2\theta\int_{t_0}^{t_1}\overline{K}d\tau. \tag{41}$$

In case there is no environmental energy then the dissipated energy vanishes as $t_1 \to \infty$, that is $\lim_{t_1\to\infty}\overline{K}(t_1) = 0$. This is an immediate consequence of (41), which could not be verified in the opposite case because of the finiteness of the initial $\overline{K}(t_0) + \overline{V}(t_0)$ and final $\overline{K}(t_1) + \overline{V}(t_1)$ energies, respectively.

[5] For the sake of simplicity, we overload the notation and use H and K instead of the corresponding $\tilde{H}$ and $\tilde{K}$.

MACHINE LEARNING

Let us consider the problem of learning introduced in Section 2. According to equation (3) the Lagrangian can be split into the kinetic term and into the loss term, which comes from the environmental constraints. Like for the case of classic mechanics, let us choose $T = D$ and $\gamma = -1$. As we have already seen, this leads to a cognitive action.[6] From (14) we get

$$\upsilon_i = \frac{\partial F}{\partial \dot{w}_i} = m_i \psi \dot{w}_i,$$

and, therefore, like for mechanics we can construct the Hamiltonian from equations (38). The function $H = K - \gamma V$ is referred to as the *energy* of the learning system. Since $\{\cdot,\cdot\}$ is linear in their first argument, we have $\{K, H\} = \{H, H\} + \gamma\{V, H\}$. Like for any learning process, even the simple case of supervised learning does require to involve the environmental potential energy $\mathcal{E}$, which arises because of the external constraints that are posed to the agent by the environment. From the general energy balance (41), that clearly holds also for learning, we can see that the environmental potential energy, supplied by the supervisor, must be dissipated. Interestingly, any learning process can be thought of as a special translation of such a transformation between the environmental and dissipated energies, with the purpose of changing the potential energy, which prescribes the effectiveness of the environmental constraints. As shown in Section 3, when considering higher-order T operators, we get more complex dynamics, as well as structural differences (see e.g. the even-order γ sign flip). In those cases, the equilibrium presents new facets that are mostly due to the extended definition of kinetic energy given in this paper, where its vanishing does not simply mean that $Dw_i = 0$ but $Tw_i = 0$. Hence there are induced dynamics at near-equilibrium points where the trajectory is in the kernel of T.

While the formulation of supervised learning involves distributional equations, in general learning mechanisms formulated in the theory we deal with ordinary differential equations that are strongly characterized by the potential energy. The re-formulation of machine learning given in [2] offers a general mechanism to construct potential energies capable of modeling also unsupervised and semi-supervised schemes.

Notice that $\mathcal{E}$ involves the environmental interactions that consist of a structural change of $V(t, w)$, but as shown in Example 5.2, other environmental interactions can arise from the bartering energy Υ, that can be reduced to temporal variations of T. Finally, it is worth mentioning that the general form of the cognitive action defined by the Lagrangian $F(t, w_i, T_i w_i)$ corresponds with recurrent neural networks in their most general non-linear form.

7. Supervised learning and collapsing of dimensionality

The theory that has been proposed gives information-based laws of learning, and it is clearly independent of the nature of the learning agent. In this section, we give insights on how the theory can originate machine learning algorithms and, in particular, we prove that the classic gradient algorithm of supervised learning can be derived from the Euler–Lagrange equations (9).

In the case of supervised learning the Euler–Lagrange equations are

$$m_i T^\star(\psi T w_i(t)) + \gamma \psi \sum_{\kappa \in \triangleright \mathcal{L}_t} (f(w(t_\kappa), u(t_\kappa)) - s_\kappa) f_{w_i}(w(t_\kappa), u(t_\kappa)) \delta(t - t_\kappa) = 0, \tag{42}$$

where $i = 1, \ldots, n$. Now we look for differential operators L such that

$$\psi L w_i = T^\star(\psi T w_i). \tag{43}$$

In the case in which $\psi(t) = e^{\theta t}$ and the coefficients α_j are time-invariant, when using Proposition 5, we have

$$\begin{aligned} T^\star(\psi T w_i) &= \sum_{h=1}^{\ell} (-1)^h \alpha_h D^h (e^{\theta t} \sum_{\kappa=1}^{\ell} \alpha_\kappa D^\kappa w_i) \\ &= \sum_{\kappa=1}^{\ell} \sum_{h=1}^{\ell} (-1)^h \alpha_h \alpha_\kappa \sum_{r=0}^{h} \binom{h}{r} \theta^r e^{\theta t} D^{h+\kappa-r} w_i \\ &= \underbrace{e^{\theta t}}_{\psi} \underbrace{\sum_{\kappa=0}^{\ell} \sum_{h=1}^{\ell} \sum_{r=1}^{h} (-1)^h \binom{h}{r} \alpha_h \alpha_\kappa \theta^r D^{h+\kappa-r} w_i}_{L w_i}. \end{aligned}$$

[6] Notice that one could reproduce the same Lagrangian by choosing the operator $T = \sqrt{\psi} D$, where $\psi(t) = e^{\theta t}$. In that case, however, as shown in Lemma 5.1 and Theorem 5.2 the boundary and bartering energies M and Υ are not necessarily null and, therefore, the energy balance needs to involve all the terms of equation (35).

Here we can see[7] that equation (43) leads to discovering L, which is a differential operator of order[8] 2ℓ.

Let us consider $t_0 = 0$ and let $g(\cdot)$ such that

$$Lg = \delta. \tag{44}$$

where $g(t) = 0$ for $t < 0$. This is a linear distributional differential equation of order 2ℓ. From the initial value theorem, we can promptly see that

$$\lim_{t\to 0} g(t) = \lim_{s\to\infty} sG(s) = 0. \tag{45}$$

Because of the linearity, we can invoke the superposition principle, we can immediately see that

$$\hat{w}_i(t) = -\gamma \sum_{\kappa \in \triangleright \mathcal{L}_t} \underbrace{\frac{f(w(t_\kappa), u(t_\kappa)) - s_\kappa}{m_i} f_{w_i}(w(t_\kappa), u(t_\kappa))}_{\zeta_{\kappa,i}} \; g(t - t_\kappa), \tag{46}$$

is a solution of (42), where $g(t) = \sum_{j=1}^{\ell} q_j e^{\rho_j t}$, and $\rho_j \in \rho(T)$ is one of the ℓ roots of the characteristic polynomial of T. Clearly, if $w_i^o(\cdot) \in \mathcal{N}(T)$ is any function which belongs to the kernel of T, then $w_i(\cdot) = \hat{w}_i(\cdot) + w_i^o(\cdot)$ is also a solution of (42). The space $\mathcal{N}(T)$ is expressed by $w_i^o(t) = \sum_{j=1}^{p} \omega_j e^{\rho_j t}$, where ω_j are uniquely defined by the initial conditions $D^r w_i^o(0)$, with $r = 0, \ldots, p-1$. In the case of in which $\Re(\rho(T)) < 0$, that is when the characteristic polynomial of T has roots with negative real part, then $\lim_{t\to\infty} |w_i(t) - \hat{w}_i(t)| = 0$. Because of equation (45), for $t = t_\kappa$ we get

$$w_i(t_\kappa) = -\gamma \sum_{h<\kappa} \zeta_{h,i} \; g(t_\kappa - t_h). \tag{47}$$

In the case of in which $\Re(\rho(T)) < 0$, that is when the characteristic polynomial of T has roots with negative real part, the computation of w according to equation (46) can be carried out by considering that $g(t_\kappa - t_h) \simeq 0$ when $t_\kappa - t_h$ is large with respect to the natural modes of the dynamics induced by T.

Now, the above equation can be nicely approximated by a recurrent form as follows[9]

$$w_i(t_{\kappa+1}) = w_i(t_\kappa) - \gamma \, g(t_{\kappa+1} - t_\kappa) \; \zeta_{\kappa,i} \tag{48}$$

Interestingly, this is the classic gradient algorithm of machine learning, where the learning rate is properly given by $\eta_{\kappa+1} = \gamma \, g(t_{\kappa+1} - t_\kappa)$. This equation opens the doors for in-depth experimentation of the algorithm with the appropriate selection of η prescribed by the theory. A preliminary experimental analysis is given in [9].

Finally, it is worth mentioning that the representation of the solution given by equation (47) strongly resembles the one of the representer theorem of kernel machines.[10] Both theory are based on differential operators; the role of kernels as the Green function of the differential operator is now played by the response impulse. The difference that one easily expects is that, because of causality, the response impulse is asymmetric ($g(t) = 0$ for $t < 0$). However, unlike in classic kernel machines, the temporal structure makes it possible to compute the weights by the recurrent equation (48).

8. Conclusion

In this paper we propose a natural learning theory that is based on the variational principle of least cognitive action. The theory allows us to understand the evolution of synaptic connections, regardless of assumptions on the structure of the agent. While there are close connections with mechanics, the given variational formulation relies on high-order differential operators with time-dependent coefficients, that give rise to richer dynamics nicely captured in the new framework of the BG-brackets. Interestingly, we show that in the case of supervised learning, the given laws collapse to the classic on gradient algorithm. The theory enlightens fundamental issues connected with the adoption of higher-order differential operators, including the even order γ sign flip. It is shown that the optimal solution is a temporal expansion of the impulse response over the training data, which is the dual of the representer theorem at the basis of kernel machines.

Finally, it is worth mentioning that all studies in machine learning based on modeling the environmental interactions by constraints are candidates for the application of the theory. In particular, this framework strongly emphasizes the role of time in the learning process, so as the evolution of the weights of the neural synapses follows equations that resemble laws of physics, driven by appropriate cognitive actions. We are currently investigating the impact of the results presented in this paper in computer vision and, particularly, in semantic labeling.

[7] Equation (43) can be satisfied also by other types of dissipation functions.

[8] From this equation we can also see the even order γ sign flip.

[9] A perfect recurrent realization requires a 2ℓ dimensional system.

[10] A formulation of kernel machines that clearly presents the duality with this theory is given in [12].

Acknowledgements[11]

We thank Duccio Papini, Paolo Nistri, Marcello Pelillo, Fülöp Baszó, Marco Maggini, Alessandro Rossi, Marcello Sanguineti, Giorgio Gnecco, and Delfim Torres for insightful discussions.

References

[1] M. Gori, M. Lippi, M. Maggini, S. Melacci, Learning to see like children: proof of concept, Tech. rep., University of Siena, 2014.
[2] G. Gnecco, M. Gori, S. Melacci, M. Sanguineti, Foundations on support constraint machines, Neural Comput. 27 (2) (2015) 388–480, http://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00686#.VKRRLYrF8Yc.
[3] M. Diligenti, M. Gori, M. Maggini, L. Rigutini, Bridging logic and kernel machines, Mach. Learn. 86 (1) (2012) 57–88, http://dx.doi.org/10.1007/s10994-011-5243-x.
[4] B. Scholkopf, A. Smola, Learning with Kernels, The MIT Press, 2002.
[5] L. Herrera, L. Nunez, A. Patino, H. Rago, A variational principle and the classical and quantum mechanics of the damped harmonic oscillator, Amer. J. Phys. 53 (3) (1985) 273.
[6] S. Frandina, M. Gori, M. Lippi, M. Maggini, S. Melacci, Variational foundations of online backpropagation, in: V. Mladenov, P.D. Koprinkova-Hristova, G. Palm, A.E.P. Villa, B. Appollini, N. Kasabov (Eds.), Artificial Neural Networks and Machine Learning, ICANN 2013, Sofia, Bulgaria, September 10–13, 2013, 23rd, in: Lecture Notes in Computer Science, vol. 8131, Springer, 2013, pp. 82–89.
[7] I. Gelfand, S. Fomin, Calculus of Variations, Dover publications, Inc., 1963.
[8] M. Giaquinta, S. Hildebrand, Calculus of Variations I, vol. 1, Springer, 1996.
[9] M. Gori, M. Maggini, A. Rossi, The principle of cognitive action – experimental analysis, Tech. rep., University of Siena, Italy, 2015.
[10] H. Bohr, Almost Periodic Functions, Ams Chelsea Publishing, 1947.
[11] M. Bianchini, M. Gori, Optimal learning in artificial neural networks: a review of theoretical results, Neurocomputing 13 (2–4) (1996) 313–346.
[12] G. Gnecco, M. Gori, M. Sanguineti, Learning with boundary conditions, Neural Comput. 25 (4) (2013) 1029–1106, http://dx.doi.org/10.1162/NECO_a_00417.

[11] When I was in high school, my physics teacher – whose name was Mr. Bader – called me down one day after physics class and said, “you look bored; I want to tell you something interesting. Then he told me something which I found absolutely fascinating, and have, since then, always found fascinating. Every time the subject comes up, I work on it.” *Richard Feynman, in physics lectures, on the principle of least action.*