Inference, Learning, and Laws of Nature

S. Frandina, M. Gori, M. Lippi, M. Maggini, S. Melacci
Department of Information Engineering and Mathematical Sciences
University of Siena, Italy
{frandina,marco,lippi,maggini,mela}@diism.unisi.it

Abstract

Although inference and learning arise traditionally from different schools of thought, in the last few years they have been framed in nice unified frameworks, in the attempt to resemble clever human decision mechanisms. In this paper, however, we support the position that a true understanding of human-based inference and learning mechanisms might arise more naturally when replacing the focus on logic and probabilistic reasoning with that of cognitive laws, in the spirit of most variational laws of Nature. To this end, we propose a strong analogy between learning from constraints and analytic mechanics, which suggests us that agents living in their own environment obey laws exactly like those of particles subjected to a force field.

1 Introduction

Inference and learning have always been the subject of curiosity and in-depth investigations in the attempt to unveil their secret and grasp their meaning. As a truly manifestation of human being, they have been studied by philosophers, logicians, psychologists, as well as by scientists in artificial intelligence. Interestingly, while inference has been early framed into logic formalisms, the process of learning has been mostly attacked by statistical approaches. Nowadays, the marriage of these methodologies has been offering nice theoretical interpretations of either inference or learning which, amongst other relevant results, leads to foundations on probabilistic reasoning. Beginning from seminal studies at the end of the Eighties (see e.g. [Pearl, 1988]), nowadays there is a huge literature in the field, from which we can see significant theoretical and experimental advances. Related studies on bridging symbolic and sub-symbolic representations in neural networks have been developing under a remarkable variety of methodologies (see e.g. the Workshop series on Neural-Symbolic Learning and Reasoning `http://www.neural-symbolic.org/`). This lack of methodological focus seems to indicate that neural symbolic integration is still looking for strongly unifying approaches, driven by solid mathematical foundations like for probabilistic reasoning. On the other side, the studies on neural symbolic integration have been opening the mind to an in-depth re-thinking of inference and learning, that is well outside the borders of probability theory.

Beginning from the biological inspiration of neural network-based intelligent agents, in this paper we support the position that a natural integration of learning and inference arises when formulating the problem within the context of intelligent agents interacting on-line with the environment under the realm of cognitive laws. As time goes by, in such an environment, the agent is giving stimuli expressed in terms of constraints amongst a set of tasks and reacts by following laws emerging from the stationary point of a functional referred to as the *cognitive action*. This is strongly inspired from analytic mechanics, where the notion of particles subjected to a force field is doomed to follow the minimization of the action functional. When considering agents embedded in their environment, we naturally associate the weights of the neural network with the coordinates of the particles, the loss related to the constraints with the potential energy, and the sum of the squares of the derivative of the weights with the kinetic energy. This leads us to study the life of the agent by means of the elegant Lagrangian and Hamiltonian frameworks. However, in order to fully grasp the above links, we need to extend these formalisms with the insurgence of dissipative processes. Basically, a newborn agent begins its life with a certain potential energy and continues by changing its parameters, thus transforming that energy into kinetic energy and by dissipating the rest. As time goes by, the velocity of the weights decreases until the agent ends into a stable configuration, where all the initial potential energy is dissipated. This results in a learning mechanism paired with an inferential process, which improves as time goes. In the next section we discuss how to bridge logic and perception, while in Section 3 we propose the laws of learning as they come out from stationary point of the cognitive action. In Section 4 we show that the weights evolve according to general energy conservation principles and, finally, in Section 5 we give a perspective view on life-long learning build up on the proposed theory.

2 Bridging logic and perception

To sketch the idea, let us consider the following example. Let $x, y \in I\!R$ and let us assume that we are given some information on functions

$$A : I\!R \to \{0, 1\}$$

$$B:\ \mathbb{R}^2 \to \{0,1\}$$

defined as follows

$$\forall x \in [0,1]: A(x) = true,\ \ \forall x \notin [0,1]:\ A(x) = false \quad (1)$$
$$\forall (x,y) \in [0,1]^2:\ B(x,y) = true,$$
$$\forall (x,y) \notin [0,1]^2:\ B(x,y) = false.$$

Now suppose that an oracle gives the intelligent agent the following piece of knowledge

$$\forall x \forall y\ A(x) \wedge A(y) \Rightarrow B(x,y), \qquad (2)$$

Let $a:\ \mathbb{R} \to [0,1]$ and $b:\ \mathbb{R}^2 \to [0,1]$ be real-valued functions associated with $A(\cdot)$ and $B(\cdot,\cdot)$, respectively. Now, suppose that an intelligent agent is living on the temporal horizon $[0, t_e]$, with $t_e > 0$. Then, we can associate the granule of knowledge (2) with

$$\mathcal{V}_1(a,b) = \int_0^{t_e} a(x(t)) \cdot a(y(t))\,(1 - b(x(t),y(t)))\,dt,$$

which needs to be as small as possible if we want to approximate (2). Now, let us assume that the agent acquires also the additional supervised pairs $\{(x_\kappa, d_\kappa^a)\}_{\kappa=1}^{\ell_a}$ and $\left\{((x_\kappa, y_\kappa), d_\kappa^b)\right\}_{\kappa=1}^{\ell_b}$ of $A(\cdot)$ and $B(\cdot,\cdot)$. They come at time $\{t_\kappa^a\}_{\kappa=1}^{\ell_a}$ and $\left\{t_\kappa^b\right\}_{\kappa=1}^{\ell_b}$, respectively. Given $c_1 > 0$ and $c_2 > 0$, the process of learning does require to control the functional

$$\mathcal{V}(a,b) := c_1 \mathcal{V}_1(a,b) + c_2 \mathcal{V}_2(a,b)$$

where

$$\begin{aligned}\mathcal{V}_2(a,b) \quad &:= \quad \int_0^{t_e} \sum_{\kappa=1}^{\ell_a} h(a(x_\kappa), d_\kappa^a) \cdot \delta(t - t_\kappa^a)\ dt \\ &+ \quad \int_0^{t_e} \sum_{\kappa=1}^{\ell_b} h(b(x_\kappa, y_\kappa), d_\kappa^b) \cdot \delta(t - t_\kappa^b)\ dt,\end{aligned}$$

and h is a loss functions (e.g. the hinge loss). This way of bridging logic and learning has been properly formalized for the case of FOL in [Diligenti *et al.*, 2012] using kernel-based representations for the functions.

Beginning from this example, let us make the assumption that functions a and b each depend on a corresponding vector of weights $w_a \in \mathbb{R}^{m_a}$ and $w_b \in \mathbb{R}^{m_b}$, respectively. Notice that this is different with respect to the kernel-based solution of [Diligenti *et al.*, 2012], but most of the concepts are the same. In order to gain a general formulation, from now on, let $x \in X \subset \mathbb{R}^n$ and $f: X \times W \to \mathbb{R}^q$ be the notation for a multitask system in which q different interacting tasks transform inputs from space X using weights in $W \in \mathbb{R}^m$. Now, we like to think of f as a neural network which, given the input $x \in X$ returns $f(x,w)$. We can promptly see that there exists V such that

$$\mathcal{V}(a,b) = \mathcal{V}(f) = \int_0^{t_e} V(w(t))dt, \qquad (3)$$

where

$$\begin{aligned}V(w(t)) \quad &:= \quad c_1 a(x(t)) \cdot b(y(t))\,(1 - b(x(t),y(t))) \\ &+ \quad c_2 \sum_{\kappa=1}^{\ell_a} h(a(x_\kappa), d_\kappa^a) \cdot \delta(t - t_\kappa^a) \\ &+ \quad c_2 \sum_{\kappa=1}^{\ell_b} h(b(x_\kappa, y_\kappa), d_\kappa^b) \cdot \delta(t - t_\kappa^b).\end{aligned}$$

Of course, no matter what constraints are presented during the agent's life, the equation (3) holds true when choosing the appropriate *potential energy*.

Links with Analytic Mechanics		
var.	mach. learn.	mechanics
w_i	weight	particle
$\dot{w}_i$	weight variation	particle velocity
V	constraint penalty	potential energy
T	temporal smoothness	kinetic energy

Table 1: Links between machine learning and analytic mechanics.

3 Lagrangian cognitive laws

We propose a unified on-line formulation of learning and inference by introducing concepts that are tightly connected with analytic mechanics (see Table 1 for a summary of connections). Interestingly, the given formulation is somewhat inspired to different ways of introducing dissipation in classic Hamiltonian systems (see e.g. [Wang and Wang, 2012; Morris, 1986; Baldiotti *et al.*, 2010; Sanjuan, 1995]). We define *cognitive action* as the functional

$$\mathcal{S} = \int_0^{t_e} \mathcal{L}_\beta\ dt = \int_0^{t_e} e^{\beta t} \mathcal{L}\ dt \qquad (4)$$

where $\beta > 0$,

$$\mathcal{L}(w) = T(w) - V(w) \qquad (5)$$

is the Lagrangian, and

$$T = \frac{1}{2} \sum_{i=1}^{m} \mu_i \dot{w}_i^2(t), \qquad (6)$$

is the *cognitive kinetic energy*, where $\mu_i > 0$ is the *cognitive mass* associated with the particle i. The learning process consists of finding

$$w = \arg \min_{w \in W} \mathcal{S}(w). \qquad (7)$$

This is a classical problem in variational calculus. Any stationary point of $\mathcal{S}$ satisfies the Euler-Lagrange equations [Giaquinta and Hildebrand, 1996]

$$\frac{d}{dt}\frac{\partial \mathcal{L}_\beta}{\partial \dot{w}_i} - \frac{\partial \mathcal{L}_\beta}{\partial w_i} = 0.$$

When considering that $\mathcal{L}_\beta = e^{\beta t}\mathcal{L}$ we get

$$\beta e^{\beta t}\frac{\partial \mathcal{L}}{\partial \dot{w}_i} + e^{\beta t}\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{w}_i} - e^{\beta t}\frac{\partial \mathcal{L}}{\partial w_i} = 0.$$

Because of the definition of the Lagrangian (5), we get

$$\ddot{w}_i + \beta \dot{w}_i + \mu_i^{-1} \, V'_{w_i} = 0, \tag{8}$$

where $i = 1, \ldots, m$. When adding Cauchy's conditions, that correspond with setting $w_i(0)$ and $\dot{w}_i(0)$, the above Lagrangian cognitive equations drive the evolution of the agent. It can be proven that when we enforce a strong dissipation (high values of β) the above equation yields the classic online Backpropagation algorithm, being $\eta_i = 1/(\beta\mu_i)$ the learning rate [Frandina *et al.*, 2013]. Interestingly, the learning rate turns out to be small for large cognitive masses, which nicely matches the intuition that large masses move slowly. In addition, as already pointed out, for this classic connection to arise, we need to use values of β that are large enough. In the next section, this is given a foundation using the Hamiltonian framework extended to dissipation.

4 A dissipative Hamiltonian framework

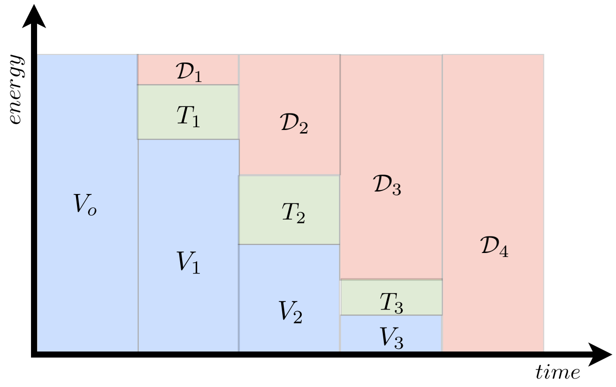

Figure 1: A sampling of energy balance. The initial value of the potential energy (inference loss) is transformed into kinetic energy and is dissipated. At the end, all the initial inference loss is dissipated.

We can start giving an interpretation of the agent evolution as follows. From equation (8) we have

$$\dot{w}_i \cdot \ddot{w}_i + \beta_i \dot{w}_i^2 + \mu_i^{-1} V'_{w_i} \cdot \dot{w}_i = 0$$

from which

$$\begin{aligned} &\sum_{i=1}^{m} \frac{1}{2}\mu_i \int_0^{t_e} \frac{d}{dt}\dot{w}_i^2 dt + \sum_{i=1}^{m} \mu_i \int_0^{t_e} \beta \dot{w}_i^2 dt \\ + &\sum_{i=1}^{m} \int_0^{t_e} V'_{w_i} \dot{w}_i dt = 0. \end{aligned}$$

Of course, we have

$$\frac{dV(w(t))}{dt} = \sum_{i=1}^{m} V'_{w_i} \dot{w}_i + \frac{\partial V}{\partial t}.$$

Then, let us define

$$D = \sum_{i=1}^{m} \beta \dot{w}_i^2 \tag{9}$$

and

$$\mathcal{D}(t) = \int_0^t D(w(\theta)) d\theta. \tag{10}$$

This is the *dissipated energy* over $[0, t_e]$. When considering definition 6, we get

$$\int_0^{t_e} \left(\frac{dT(w(t))}{dt} + \frac{dV(w(t))}{dt} + \frac{d\mathcal{D}(t)}{dt} \right) dt = \int_0^{t_e} \frac{\partial V}{\partial t} dt.$$

In case $\frac{\partial V}{\partial t} = 0$ we end up into the following *principle of conservation of cognitive energy*

$$\frac{d(T + V + \mathcal{D})}{dt} = 0 \tag{11}$$

that is the *cognitive energy*

$$\mathcal{E} = T + V + \mathcal{D}$$

is conserved during the agent's life. In Fig. 1 we can quickly grasp the meaning of the energy invariance. Basically, as time goes by, the potential V is partly transformed into kinetic energy and is partly dissipated. It is easy to see that the kinetic energy vanishes as $t_e \to \infty$. If, by contradiction $\lim_{t_e \to \infty} T(w(t)) > K > 0$ then, there exists $\bar{t}$ such that

$$\begin{aligned} \mathcal{D} &= \int_0^{t_e} \beta \dot{w}^2 dt > \int_{\bar{t}}^{t_e} \beta \dot{w}^2 dt \\ &> K(t_e - \bar{t}) \end{aligned}$$

and, therefore $\lim_{t_e \to \infty} \mathcal{D} = \infty$, from which we conclude that the kinetic energy vanishes necessarily. If the potential energy (loss term) is time-dependent (injection of stimulus) then we have the more general conservation law

$$\frac{d\mathcal{E}}{dt} = \frac{\partial V}{\partial t}. \tag{12}$$

This tells us that the cognitive energy $\mathcal{E}$ is constant whenever there is no injection of stimuli, but as the agent reacts to a new constraint there is change of cognitive energy, which is partially dissipated and transformed into kinetic energy.

INFERENCE

The proposed framework places learning and inference in exactly the same framework. In the example given in Section 2, the given constraints are properly expressed by the corresponding potential energy and the weights (particle position) evolves following the general conservation principle of equation (12). Now, suppose that we communicate to the agent the additional granule of knowledge

$$\forall x, \forall y : \; C(x) \wedge C(y) \Rightarrow D(x, y)$$

and that we want the agent to infer the truth of predicate

$$(\neg B(x, y) \wedge D(x, y)) \vee (B(x, y) \wedge \neg D(x, y)) \Rightarrow A(x).$$

We can convert this predicate to the correspondent real-valued function as shown in Section 2 and check it accordingly [1]. Of course, while some constraints are given and

[1] You can perform this kind of learning and inference on batch-mode using the simulator at `https://sites.google.com/site/semanticbasedregularization/home/software`.

are responsible of the learning dynamics (12), others are only used for inference. This is somewhat coherent with the scheme proposed in ([Diligenti *et al.*, 2012]). A detailed description of the inferential mechanisms is shown in [Gori and Melacci, 2013], even though kernel based representations are used instead of neural-like models like those considered in this paper. However, the most remarkable difference is that the learning process described in this paper takes place fully on-line. This way of inferring new constraints is different to nowadays approaches of collective classification, since the inference relies on the developed weights exactly like in classic on-line backpropagation.

5 Discussion

The cognitive laws formulated in this paper can be thought of as a re-statement of variational approaches to learning, like the one recently proposed in [Gnecco *et al.*, 2013]. There is, however, a fundamental difference that involves the crucial role of time, which is in fact what gives rise to the unified on-line learning and inferential scheme. Now, as already pointed out, the stationary points of (8) are those for which $\nabla_w V = 0$, which clearly indicates the limitations of cognitive systems that are stressing the dissipation. As pointed out for the case of supervised learning, large values of β lead to a stochastic gradient descent, but this property holds in general. While this is a nice landing in a known planet, there is still the problem of local minima of the potential energy. Interestingly, equation (8) is a damping oscillator, which can get rid of suboptimal minima energy configuration when the dissipation parameter β is small. An additional role is played by the injection of noise, which leads to the Langevin equation. In particular, there is a nice duality with the dynamics of Brownian particles, which are active in the sense that they take up energy from the environment [Schweitzer, 2000; Ebeling and Schweitzer, 2001]. The interpretation by the Fokker equation might be very interesting to gain a probabilistic understanding of the learning process. Alternative intriguing connections arise when invoking the extension of the Cognitive Action as defined in this paper in the quantum framework as early pointed by R. Feyman in his Ph.D. thesis [Feyman, 1942]. Finally, regardless of the development of the idea sketched in this paper, the remarkable novelty is in the way we seek for optimal cognitive configurations by the proposed on-line scheme, which can be thought of as laws of Nature. While there is a weak connection with simulated annealing and related global optimization schemes, the equations (8) are truly embedded in the environment and, therefore, they lead naturally to on-line learning. This seems to fit the growing call for life-long learning models, where a truly intelligent agent should be capable of learning online from a lifetime of raw sensorimotor experience (see the AAAI-2011 Workshop on lifelong learning from sensorimotor experience).

Acknowledgments

The research has been supported under the PRIN 2009 grant "Learning Techniques in Relational Domains and Their Applications".

References

[Baldiotti *et al.*, 2010] M.C. Baldiotti, R. Fresneda, and D.M. Gitman. Quantization of the damped harmonic oscillator revisited. Technical report, Istituto de Fisica, Universidade de Sao Paulo, Caixa Postal 66318-CEP, 05315-970 Sao Paul, S.P., Brasil, 2010.

[Diligenti *et al.*, 2012] M. Diligenti, M. Gori, M. Maggini, and L. Rigutini. Bridging logic and kernel machines. *Machine Learning*, 86:57–88, 2012. 10.1007/s10994-011-5243-x.

[Ebeling and Schweitzer, 2001] W. Ebeling and F. Schweitzer. Active motion in systems with energy supply. In *Integrative Systems Approaches to Natural and Social Dynamics*, pages 119–142, Berlin, 2001. Springer.

[Feyman, 1942] R. P. Feyman. *Principles of Least Action in Quantum Mechanics*. Ph.D. thesis, Princeton University, 1942.

[Frandina *et al.*, 2013] S. Frandina, M Gori, M Lippi, M Maggini, and S. Melacci. Variational foundations of online backpropagation. Technical report, Department of Information Engineering and Mathematical Sciences, University of Siena, 2013.

[Giaquinta and Hildebrand, 1996] M. Giaquinta and S. Hildebrand. *Calculus of Variations I*, volume 1. Springer, 1996.

[Gnecco *et al.*, 2013] G. Gnecco, M. Gori, and M. Sanguineti. Learning with boundary conditions. *Neural Computation*, pages 1–78, 2013.

[Gori and Melacci, 2013] M. Gori and S. Melacci. Constraint verification with kernel machines. *IEEE Trans. on Neural Networks and Learning Systems*, 24(5):825–831, 2013.

[Morris, 1986] Philip J. Morris. A paradigm for joint hamilton and dissipative systems. *Physica D*, pages 410–419, 1986.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffmann, San Francisco, California, 1988.

[Sanjuan, 1995] M.A.F. Sanjuan. Comments on the hamiltonian formulation for linear and nonlinear oscillators including dissipation. *Journal of Sound and Vibration*, 185 (4):734–736, 1995.

[Schweitzer, 2000] F. Schweitzer. Active motion of brownian particles. In *Stochastic Processes in Physics, Chemistry and Biology*, pages 97–106, Berlin, 2000. Springer, Lecture Notes in Physics, vol. 557.

[Wang and Wang, 2012] Q.A. Wang and R. Wang. Is it possible to formulate least action principle for dissipative systems? Technical report, arXiv, 2012.