UNIVERSITÀ
DEGLI STUDI
FIRENZE

PHD PROGRAM IN SMART COMPUTING
DIPARTIMENTO DI INGEGNERIA DELLINFORMAZIONE (DINFO)

A VARIATIONAL FRAMEWORK FOR LAWS OF LEARNING

ALESSANDRO BETTI

Dissertation presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Smart Computing

This page intentionally left blank

PhD Program in Smart Computing
University of Florence, University of Pisa, University of Siena

A VARIATIONAL FRAMEWORK FOR LAWS OF LEARNING

ALESSANDRO BETTI

Advisor: **Prof. MARCO GORI**

Head of the PhD Program: **Prof. PAOLO FRASCONI**

Evaluation Committee:
Prof. MICHAEL BRONSTEIN *Imperial College London*
Prof. ULISSE STEFANELLI *University of Vienna*

XXXII ciclo — October 2019

I imagine this midnight moments forest:
Something else is alive
Beside the clocks loneliness
And this blank page where my fingers move.

— TED HUGHES, *The Hawk in the Rain* (1957)

To my grandfather Ettore,
whose love for knowledge
has been an absolute model
and source of inspiration.

Acknowledgments. I would like to thank my advisor Marco Gori for his guidance throughout my PhD studies and for all the long and invaluable discussions we had during these years. His constant presence and continuous interaction has been the main source to many of the ideas at the basis of this work. I thank Stefano Melacci for his crucial contribution to the development of software and to the preparation and conduction of the experimental campaigns. A very special thank goes to Giovanni Bellettini who assisted and advised me on many mathematical issues mainly concerning variational calculus. Many thanks to the members of my Supervisory Committee, Giuseppe Boccignone and Stefano Soatto for their authoritative feedback on my work.

PREFACE

Set N equal to 1
Begin reading Chapter N.
— DONALD E. KNUTH, The Art of Computer Programming vol. 1 (1997)

As Galileo expressed very plainly in its famous work *Il Saggiatore*, the way which has turned out to be most effective to describe natural phenomena is though the study of mathematical models which are able to capture, to some extent, the salient features (or the ones that we believe to be mostly relevant) of the process that we observe. He stated indeed that:

> *La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.*

This *translation* into mathematical language is usually done through a mathematical framework that is expressive enough to host all the possible configurations of interest. To do this we typically end up with spaces of configurations that are much richer that the one that we can observe. Then the ability of the translator (the scientist) is to find relations and constraints between the parameters that defines the large space of configurations that render the behaviour of the model similar to the behaviour observed in nature. Scientists usually call this set of relations *laws*.

A very important type of laws are evolution laws; that is to say the rules by which the configurations of our mathematical description changes in time when we start from a known configuration. For example Newton's laws of motion are evolution laws, while, on the other hand, the laws that characterizes the shape of a chain that hangs on a wall nailed to it by the two endpoints is not, as it is concerned with the relative position of the links of the chain against the wall. However physics has showed us many times how most of the phenomena that we perceive as stationary are actually the asymptotic value of a dynamical process and that the knowledge of the laws of evolution in most of the cases is much more informative than the knowledge of only the laws of a system at equilibrium.

The aim of this work is to start to investigate whether and how it is possible to formulate learning as an evolution problem.

The basic intuition behind this idea is the fact that many problems in learning naturally presents themselves as a coherent stream of information which have its proper dynamics and temporal scales; one emblematic example is that of visual information. However nowadays most of the approaches to learning completely disregard, at all, or in first approximation, this property of the information on which the learning should be performed. As a result the problem is typically formulated as a "static" optimization problem on the parameters that define a learning model.

Formulating a learning theory in terms of evolution laws instead shifts the attention to the dynamical behaviour of the learner during the process of learning. As we commented above this gives us the opportunity, for those agents that lives into streams of data, to couple their dynamics with the information that flows from the environment. Another advantage that we see in this approach is the possibility to incorporate into the temporal laws of learning dynamical constraints that, for example, we know that they will enhance the quality of the leaning.

We begin our analysis in Chapter 1 by showing how typical machine learning problems can be defined when data is available as a function of time. In particular we show the affinity between typical loss functions of ML and potential functions of classical mechanics both in the time independent case (what in machine learning we would call batch mode) and the time dependent case (that will induce an online learning problem). Then we go on and discuss possible temporal evolutions laws for the parameters that defines the learning model. In the last section of the chapter we give an important consistency results for a class of evolution laws that shows how, in the case where there is interaction with an environment, the out-of-equilibrium dynamics gives consistent predictions on consistent external stimuli.

The analysis of the first chapter is carried on in Chapter 2 where we show how we can formulate learning theories using variational methods. Despite the natural way in which learning problem can be formulated by means of calculus of variations we faced and analyzed the problem of causality that very soon rises when we try to formalize an evolution problem using integral functional index. Eventually we will apply the theory of lagrange multipliers for variational problems to show how it is possible to reformulate learning problems concerning NN in terms of constrained variational problems; this general theory is particularly useful in view of the material discussed in the last chapter.

Finally Chapter 3 is devoted to the development of a theory of extraction of visual features from videos mostly using unsupervised techniques. The whole theory is based on three different kind of principle:

1. Regularity;
2. Maximization of mutual information;
3. Enforcement of a dynamical consistency constraint.

These principles are then used to define a variational index based on convolutional features from which the dynamics of the convolutional filters are derived.

This last Chapter closes with a section that revises some aspects of the theory proposed in the first two sections of the chapter on the basis of experimental results and on the analyses described in Chapter 1 and 2.

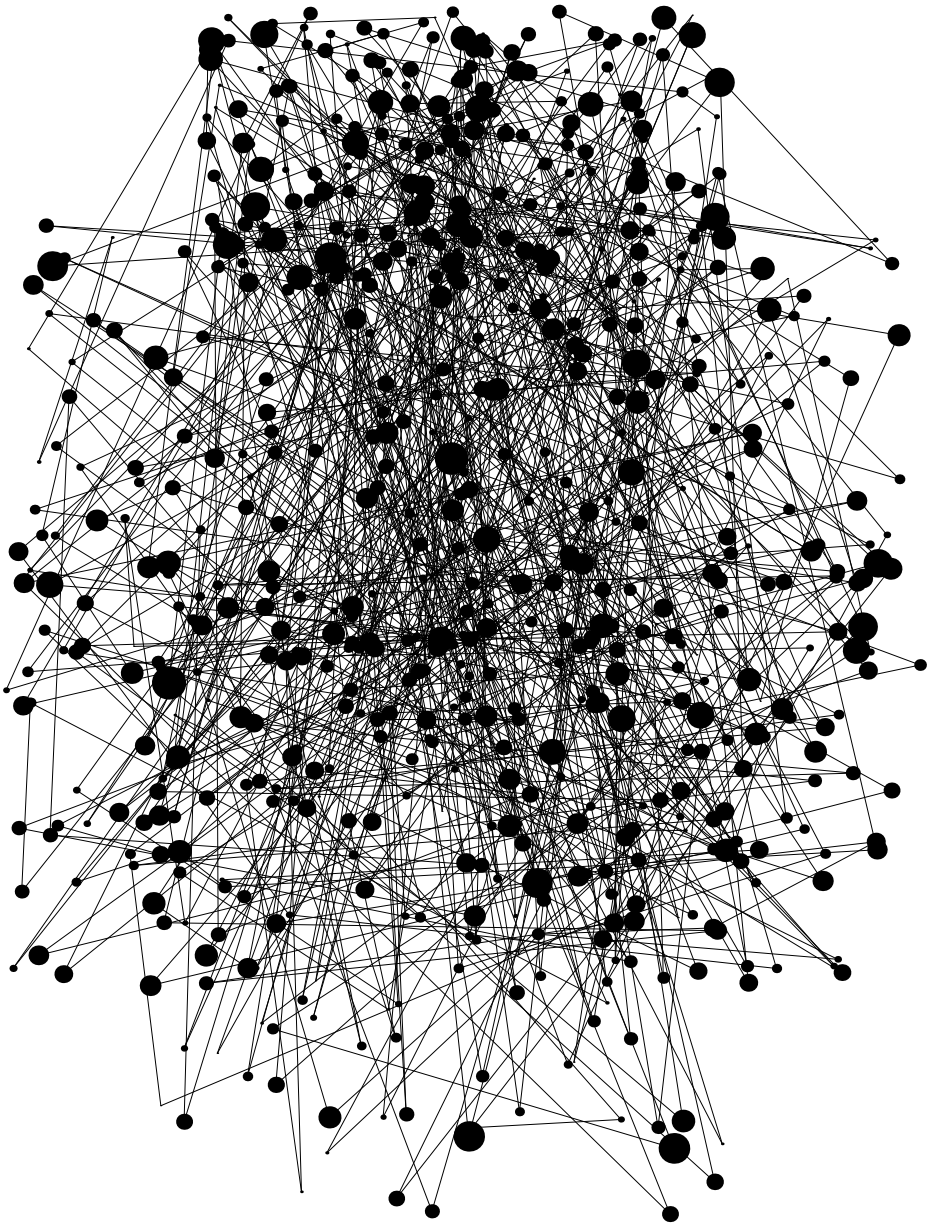
Firenze A.B.
October 2019

In the process of learning to cook
we build an intuition about
the underlying science as well.
We know that a copper pan heats more evenly
than one made with iron, although we may
not be able to explain why.

— Modernist Cuisine, vol. 1, (2011)

CONTENTS

Thou art the Mars of malcontents.
— WILLIAM SHAKESPEARE, *The Merry Wives of Windsor* (I, iii, 111)



CHAPTER ONE

LEARNING IN ERGODIC ENVIRONMENTS

Imagine how hard it would be to use a dictionary if its words were not alphabetized!

— DONALD E. KNUTH, in *The Art of Computer Programming, vol. 3, Sorting and Searching* (1998)

Time as he grows old teaches all things.

— AESCHYLUS, in *Prometheus Bound* (l. 981)

All kinds of learning are concerned with data; more precisely learning consists in the exposure and the interaction of a learning agent (an entity that has some degree of adaptation) with information that comes from the environment. In machine learning agents are computers while environments consists of sets of examples that are typically static structures in which the program that implements the learning algorithm have free access and, in principle, can retrieve any kind of information at any step of the process of learning.

This scenario, of course, contrasts with the way in which human agents interacts with data. The environment in which humans dwell is not static, and the agents do not have direct access to any kind of information any time they want: Memory indeed is a property of the agent not of the environment itself. What we can observe is that humans (or any other known biological agent with learning capabilities) live and learn in temporal environments.

In this chapter we will discuss how it is possible to reformulate classical concepts of machine learning when we work with temporal signals (one dimensional manifolds) in an Euclidean space: $t \mapsto x(t) \in \Omega$, with Ω generally being a subset of a high dimensional euclidean space that contains all the learnable information.

The duality that can be established between a probability space (like the one in which ML is usually defined) and suitable trajectories on the space itself that relates the probability associated with a small volume of the space with the time spent by the trajectory in the same volume, is a fundamental concept used fruitfully both in physics and, more generally, in the study of probability spaces. In statistical mechanics Bolzmann [8] introduced the ergodic hypothesis in order to have such relation for trajectories in phase space; more precisely Bolzmann hypothesized that each orbit starting from an initial point in phase space having a certain energy would generate the the whole energy surface (indeed the word ergodic is a composite word that originates from the Greek words ἔργον that means *work* and ὁδός that is *path*). From this hypothesis Bolzmann, showed

that means in time and in phase space are equal. Nowadays we usually refers to ergodicity as the properties that the flow in phase space needs to have in order to have equal means in space and in time over real valued functions (see [57]). More generally ergodic theory now study the actions of groups on measure spaces.

In statistical mechanics, Bolzmann ideas were used to transform time averages of quantities evaluated on huge numbers (of the order of magnitude of the Avogadro number) of trajectories described by two times many first-order differential equations (Hamilton's equations) in averages over phase space with a measure that for some thermodynamical systems (see [29] for precise statements) can be derived explicitly. In this sense we can say that statistical mechanics takes advantage of ergodicity to go from time formulation to timeless formulation. On the opposite, all the derivations presented in this chapter are based on the belief that that in (some problems of) learning is exactly the other way around.

We will begin by showing the ergodic translation of the functional risk in machine learning; this will represent the starting point to reformulate learning as a class of problems directly in the temporal dimension and to discuss and analyze different possible continuous temporal dynamics for the parameters of a learning model.

1.1. ERGODIC ASSUMPTION IN LEARNING

Because of the generality of the arguments that we will be discussing below we do not need to focus on a particular type of machine learning problem; indeed it is not even necessary to distinguish (up to a certain point) between supervised and unsupervised problems. We only require to have the following spaces and structures:

1. An environmental space Ω that contains all the information relevant for the learning;
2. A learning machine (sometimes also called agent) characterized by some adjustable parameters $\omega \in \mathbf{R}^N$;
3. A loss function v that measure how much the agent is "learning" from the environment.

We shall also assume that the information that comes from the environment can always be represented as elements of a finite dimensional real vector space.

In standard (statistical) learning theory [56] it is assumed that the data in Ω are generated according to some fixed probability distribution function, so that the setting in which the theory is framed is that of measure theory. We will now show how this ideas together with ergodicity naturally suggests a transition from risk functions to potentials.

1.1.1. Risks and Potentials

First of all assume, as already mentioned, that the environment information is encoded into points of

$$\Omega \subset \begin{cases} \mathbf{R}^d \times \mathbf{R}^n, & \text{supervised learning;} \\ \mathbf{R}^d, & \text{unsupervised learning.} \end{cases} \tag{1}$$

Indeed in supervised learning the environment both supplies the examples (points in $\mathbf{R}^d$) and the supervisions (points of $\mathbf{R}^n$) while unsupervised tasks are all about data. One typically thinks that the learner is exposed to the information of Ω drawn with a specific distribution that completely characterize the ML task that one is facing. We let $\pi\colon \mathcal{F} \subset \wp(\Omega) \to [0,1]$, $\pi(\Omega) = 1$ be a probability measure over the set Ω; here $\wp(\Omega)$ is the power set of Ω. Since we are working with subsets of real vector spaces we mainly think to π as strictly related to the Lebesgue measure.

The other missing ingredient is a way of measuring the performances of the agent; this can be done by means of a loss function $v\colon \mathbf{R}^N \times \Omega \to [0, +\infty]$ that we assume to be $v(\omega, \cdot)$ π-integrable. Such function takes as inputs the parameters ω that describe the agent in a certain stage of learning and an example from the environment and establish how appropriate are the weights* ω for the task that we want to learn**.

* In neural network the parameters are the weights of the neurons. For this reason in what follows we will use the word "weights" as a synonymous of parameters even when we are not considering a NN.

** Sometimes in the ML literature (see for example [21]), the loss function is considered to be a function that takes as an input not just the model parameters ω and the data point $x \in \Omega$

The risk functional (or functional risk) is classically defined as: Take $X\colon \Omega \to \Omega$ to be the identity random variable and consider

$$V(\omega) := \mathrm{E}\, v(\omega, X) = \int_\Omega v(\omega, x)\, d\pi(x). \tag{2}$$

The typical problem of machine learning can be then concisely summed up as the search of the "best" parameters w in the sense that

$$w = \arg\min\{V(\omega) \mid \omega \in \mathbf{R}^N\}. \tag{3}$$

Since the probability measure π is not usually known, a standard procedure is that of approximating V with the *empirical risk*: Suppose that instead of π you get to know a sample of ℓ elements $X_1 = x_1, \ldots, X_\ell = x_\ell$. Then a very crude (but effective) way of approximating π is the following:

$$\pi \approx \pi_\ell := \frac{1}{\ell} \sum_{k=1}^{\ell} \delta_{x_k}, \tag{4}$$

where* $\delta_p(T) := [p \in T]$ for all $T \subset \Omega$ is the Dirac measure. In doing so the expectation value in (2) becomes a sum:

$$V(\omega) \approx \int_\Omega v(\omega, x)\, d\pi_\ell(x) = \frac{1}{\ell} \sum_{k=1}^{\ell} v(\omega, x_k), \tag{5}$$

that is what is usually referred to as empirical risk [56].

As anticipated at the beginning of this chapter, however, we are interested into another kind of representation of the risk functional, namely one based on the existence of a trajectory $t \mapsto x(t)$ in the environment space Ω. For the moment we do not assume any kind of regularity on the function x other than that of being Lebesgue measurable. On the other hand we make a strong ergodic hypothesis on x, namely we assume that x is such that

$$U(\omega) := \lim_{T\to\infty} \frac{1}{T} \int_0^T v(\omega, x(t))\, dt = V(\omega). \tag{6}$$

In general this can be quite a demanding hypothesis on the form of x and one may rise concerns even on the existence of such trajectories. Therefore some observations are due at this point. First of all it is important to notice that as it happens for the empirical risk, also Eq. (6) can be in principle rather effective for learning purposes also if it holds only up to a certain approximation. The second, and most pertinent observation is the fact that for many learning tasks (like many of the tasks that human can solve in vision) we believe that the relevant statistic is the one contained in a temporal signal and, as it happens in physics, the probability distribution that appears in the risk functional can be

but these quantities are given to the criterion trough the model itself: If $f(\omega, x)$ is the model for computing the output then the loss is sometimes defined as $v(f(\omega, x), x)$.

* We use the Iverson's bracket notation [statement]; this quantity is equal to 1 if statement is true, and to 0 otherwise.

constructed from it. Elaborating on this idea, of course, all the tasks in which the probability measure is constructed starting from a temporal signal by definition clearly satisfy (6).

Relation (6) is a first step towards a possible reformulation of machine learning where temporal dynamics of the parameters of the model are considered from a principled way. The strict correspondence that we get here is restricted to the case when the environmental information is, in some sense, accumulated in a time independent quantity*. However one may start to ask what happens if we start to define a local-in-time criterion for the parameters of the model based on the environmental information.

At this point in order to avoid confusion in talking about temporal dynamics it is necessary to distinguish between two important and separate concepts:

1. The temporal dynamics of the environment, that is defined by $x(t)$;
2. The dynamics of learning that basically it is the way in which the parameters of the underlying model change to adapt and learn from the environment.

Of course 2. is more general than 1. in the sense that dynamics of the parameters (as we will see for example in batch mode learning) can be based on mechanisms that do not require a *temporal interaction* with the environment, while the knowledge of the evolution processes of the environment in itself, if not related to the dynamics of the parameters, is of no interest in learning. It is also natural to think whether it is always better to formulate learning problems in terms of the optimization problem stated in (3) or if it may be the case of looking for proper *laws of learning* [5] which prescribe the whole evolution of weights possibly taking into account possible stimuli that comes from the environment and, of course, go eventually in the direction of improving learning performances.

In order to stress this change of prospective, from learning algorithms crucially based on optimization techniques that are mere tools to accomplish (3) to implementations of proper laws of learning, we would also like to replace the term risk with the more evocative *potential*. In doing so we also legitimate explicit temporal (in the sense of 1. above) dependencies of the potential.

We will now proceed at examining possible ways to construct potentials starting from the loss function; then we will begin to look for possible the dynamical laws for the parameters based on the potentials.

Learning Potentials. The simplest and ready-to-use potential that ensure direct connections not only with classical statistical learning but also with classical mechanics is the *batch potential* $U(\omega)$ as it is defined in Eq. (6). This is a time independent potential that, under the ergodic hypothesis, coincides with the risk $V(\omega)$. Whenever ergodic hypothesis does not hold, $U(\omega)$ can still be thought of as an approximation of the empirical risk.

* In Eq. (6) after all we are still considering averages so that we cannot tell the differences in learning if we consider the trajectory with opposite velocity in which the environment is visited in the reversed order

In the opposite direction goes the fully "online" potential

$$U(\omega, t) := v(\omega, x(t)), \tag{7}$$

which has an explicit time dependence that describes the interaction with the environment. Notice that for constant signals this potential collapses, as it should, to the batch potential so we may conjecture that for slowly varying signals online and batch potentials may give similar learning phenomena.

In between the two extreme cases, we can also came up with some time-dependent potential that bears memory of the past information:

$$U(\omega, t) := \frac{1}{t} \int_0^t v(\omega, x(s))\, ds. \tag{8}$$

For the moment let us consider these three kind of potentials and let us now focus on the definition of the learning problem itself. Later in this chapter we will analyze another potential that can be regarded as a natural transposition in time of the regularization theory proposed by T. Poggio and F. Girosi [44].

1.2. LAWS OF MOTION

Driven by the idea that we can model a learning process in terms of interaction with the environment, we allowed our parameter selection criterion to have an explicit time dependence. In terms of the definition of the learning problem this has, of course, some repercussions. While batch mode learning can be always be formulated by means of Eq. 1.1.1–(3), when using a potential that explicitly depends on time it is no more clear how the problem should be formulated. The first guess is that one can define the evolution of the parameters, i.e. construct the map $t \mapsto w(t)$, by solving a series of minima problems

$$w(t) := \arg\min_{\omega \in \mathbf{R}^N} U(\omega, t), \tag{1}$$

one for each instant of time t.

However this first attempt does not yield an interesting learning behaviour as for each t we are solving an independent problem. For example if the dependence on t is periodic (assuming all the minimum problems are well defined) the resulting dynamics for w would be an oscillatory non-convergent dynamics.

The simplest solution for this problem is to define a dynamics that *updates* the weights based on a goal, the potential — possibly time variant — and to the previous values of the weights. A very simple mathematical model that nicely interprets this idea, is an ordinary differential equation (ODE). So we will henceforth assume that the trajectory $t \mapsto w(t)$ is the solution of an ODE: After all, the most used optimization method for 1.1.1–(3) used in ML is gradient descent (GD) which is nothing but the discrete version of a first order ODE. In the remaining part of this section we will analyze the dynamics yield by a first-order and second-order ODEs. The structure of these equations will be inspired to that of classical mechanics, the idea being that mechanical systems, when they are allowed to loose energy, will try to sit into a minimum of the potential.

1.2.1. Time-independent Potentials

In order to easily compare this with the problem stated in 1.1.1–(3) we will, for now, restrict ourselves to the time-independent case. In the next section we will instead focus on time-dependent potentials.

First order equation. The first first-order dynamics that we will consider is a gradient flow dynamic:

$$\dot{w}(t) = -\eta \nabla U(w(t)), \quad \eta > 0, \tag{1}$$

where $\nabla U(\omega)$ is the gradient of U. In terms of learning this, as we already mentioned, is the analogous of full GD on the potential U, which is the standard ML method with which the minimization of the empirical risk is attained.

The physical interpretation of these equations is not completely apparent; however these where the equation of motion for particles proposed by Aristotele* [Ph 215a25].

* Actually from Aristotele's arguments one can only deduce that the speed of a body (in a gravitational field) would go to infinity when we let the density of the fluid go to zero. Recently,

This simple ODE has very nice relations with the minimum problem 1.1.1–(3) since it represent a gradient flow on the function U. Indeed by multiplying Eq. (1) from both sides by $\dot{w}$ we get $|w(t)|^2 = -\eta \nabla U(w(t)) \cdot \dot{w}(t)$ that implies

$$\frac{dU(w(t))}{dt} = -\frac{1}{\eta}|\dot{w}(t)|^2 < 0. \tag{2}$$

This is the statement that U is a Lyapunov function.

Second order equation. The analogous of Eq. (1) is

$$\mu \ddot{w}(t) = -\nabla U(w(t)). \tag{3}$$

This equation however, as it is well known, has a very different behaviour. By again multiplying each side of Eq. (3) by $\dot{w}$ we get

$$\frac{d}{dt}U(w(t)) = -\frac{d}{dt}\left(\frac{1}{2}|\dot{w}(t)|^2\right), \tag{4}$$

so that when the magnitude of the velocity $\dot{w}$ goes down the potential goes up and vice versa; indeed we know that Eq. (3) admits oscillatory solutions. This being said, this model, that in physics represents the motion of a particle subject only to conservative forces, is not of interest in learning.

Consider instead the following equation

$$\mu \ddot{w}(t) + \theta \dot{w}(t) + \nabla U(w(t)) = 0 \qquad \mu, \theta > 0. \tag{5}$$

Notice that Eq. (5) is the continuous form of a classic multistep first-order method (see [46]):

$$w^{k+1} = w^k - \alpha \nabla U(w^k) + \beta(w^k - w^{k-1}), \qquad \alpha, \beta \geq 0, \tag{6}$$

called the *heavy ball method* since Eq. (5) can be interpreted as the equation of motion of an heavy ball with friction subject to the potential $U(w)$. Therefore from the point of view of machine learning this dynamics really goes in the direction of minimizing the potential.

Equation (5) describes the motion of a particle that moves into a potential U subject to a drag force proportional to its velocity through the constant θ. Whenever $\mu/\theta \ll 1$ Eq. (5) is an approximation of (1) with $1/\theta \equiv \eta$.

Therefore we saw in that in the case of time independent potentials the laws of classical mechanics gives a sound method of learning that generalizes the classical GD method.

1.2.2. Time-dependent Potentials

We are now in position of attacking the problem of learning formulated as a direct interaction with a temporal environment. For this reason we consider a time dependent potential of the form $U(\omega, t)$. In what follows we will denote

in an interesting article [48] Rovelli argued that Aristotele's theory is an approximation of Newton's theory when we consider objects in spherically symmetric gravitational field that moves in a fluid.

both with ∇U and U_ω the gradient with respect to omega and with U_τ the derivative of the potential with respect to its second argument (time).

In this case the analogous of Eq. 1.2.1–(1) is, of course,

$$\dot{w}(t) = -\eta U_\omega(w(t), t). \tag{1}$$

Because of the explicit temporal dependence, the solution cannot be interpreted as a gradient flow of any function. In particular if we choose U to be as in 1.1.1–(7) we have that $w(t)$ follows the dynamics

$$\dot{w}(t) = -\eta v_\omega(w(t), x(t)), \tag{2}$$

which is the continuum analogue of a SGD method. It is well-known [9] that SGD convergence is guaranteed under strict hypothesis on η, v and the nature of $x(t)$, which are not so easy to control. However, in practice, it is known that with a learning rate η sufficiently small and if the examples $x(t)$ are properly repeated over learning then convergence is attained.

For the moment we will not pursue an in-depth analysis of this equation, instead, we will consider its second-order generalization:

$$\mu\ddot{w}(t) + \theta\dot{w}(t) + \nabla U(w(t), t) = 0 \qquad \mu, \theta > 0. \tag{3}$$

As usual (1) is recovered when $\mu/\theta \ll 1$.

Distributional Potential. A particularly interesting time-dependent potential is the following distribution

$$U(\omega, t) = \sum_{k=0}^{+\infty} v(\omega, t)\delta_{t_k}(t), \tag{4}$$

where $\langle t_n \rangle$ is a sequence of times. This particular form of time dependence is of interest both because it transposes the ideas of Poggio and Girosi described in [44] in the temporal domain and also because it allows an explicit analysis of the solutions of equation (3). More precisely consider the equation

$$\ddot{w} + \theta\dot{w} + \sum_{k=0}^{+\infty} v_\omega(w(t), t)\delta_{t_k}(t) = 0, \tag{5}$$

and take the Laplace transform of this expression*. Using the properties of Laplace transform, we get

$$s^2\mathbf{L}w(s) - sw(0) - \dot{w}(0) + \theta(s\mathbf{L}w(s) - w(0)) + \sum_{k=0}^{+\infty} v_\omega(w(t_k), t_k)e^{-st_k}. \tag{6}$$

Let us begin to study what happens with just one impulse at $t = \tau$; so consider the equation

$$\ddot{w} + \theta\dot{w} = \Phi(w)\delta_\tau, \tag{7}$$

* Remember that for a function f the Laplace transform $\mathbf{L}f(s)$ is $\int_0^{+\infty} e^{-st} f(t)\, dt$.

along with some initial conditions $w(0) = w_0$ and $\dot{w}(0) = v_0$. Now take the Laplace transform of this, so that we get

$$s^2\mathbf{L}w(s) - sw(0) - \dot{w}(0) + \theta(s\mathbf{L}w(s) - w(0)) = \Phi(w(\tau))e^{-s\tau}, \tag{8}$$

meaning that

$$\mathbf{L}w(s) = \frac{\Phi(w(\tau))e^{-s\tau} + v_0 + \theta w_0 + sw_0}{s^2 + \theta s}. \tag{9}$$

Inverting the Laplace transform we get

$$w(t) = w_0 H(t) + \frac{v_0}{\theta}(1 - e^{-\theta t})H(t) + \frac{\Phi(w(\tau))}{\theta}(1 - e^{-\theta(t-\tau)})H(t-\tau), \tag{10}$$

where H is the Heaviside function; for $t > 0$ we simply have

$$\begin{aligned} w(t) &= w_0 + \frac{v_0}{\theta}(1 - e^{-\theta t}) + \frac{\Phi(w(\tau))}{\theta}(1 - e^{-\theta(t-\tau)})H(t-\tau) \\ &= w_0 + \frac{v_0}{\theta}(1 - e^{-\theta t}) + \Phi(w(\tau))G(t-\tau), \end{aligned} \tag{11}$$

where we defined $G(t) := \theta^{-1}(1 - e^{-\theta t})H(t)$ *. If we now look back to Eq. (6) and we make use of the superposition principle we get

$$w(t) = w_0 + \frac{v_0}{\theta}(1 - e^{-\theta t}) - \sum_{k=0}^{+\infty} v_\omega(w(t_k), t_k)G(t - t_k). \tag{12}$$

Now if we define $K_t := \sup\{m \mid t_m < t\}$ this expression can be rewritten as

$$w(t) = w_0 + \frac{v_0}{\theta}(1 - e^{-\theta t}) - \sum_{k=0}^{K_t} v_\omega(w(t_k), t_k)G(t - t_k). \tag{13}$$

which shows the causality nature of the solution.

This method was originally proposed in [5] and, later on, it was developed and tested in details in [23], where it is also showed how it can be made equivalent to GD. It represent a first step in the study of systems with an explicit interaction with the environment.

---

* The function G should actually be thought as the Green function of the differential operator $d^2/dt^2 + \theta d/dt$

1.2.3. Quasi-Periodic Environments

We now turn to the general second order equation 1.2.2–(3) in order to show that this dynamic becomes consistent with learning, in the sense that after some time it will give consistent predictions on similar patterns, as soon as we start to make assumptions on the regularity of x. For the sake of simplicity we will take in Eq. 1.2.2–(3) $\mu = 1$.

Regardless of the specific probability distribution that characterizes the environmental data, when focussing on a generic sample $x(t)$, one can reasonable expect that a similar pattern will likely appear in the future. Moreover, it makes sense to assume that such a property takes place uniformly in the temporal domain, and that for any $t \in (0, \infty)$, at least one pattern in the future, at time $\overline{t}$, be similar, in the metric sense, so as $\|x(t) - x(\overline{t})\|$ is small. This idea will be at the base of the definition of quasi-periodicity.

Before introducing the notion of quasi-periodicity, we need some preliminaries.

Suppose $w(t) \in \mathbf{R}^n$ and $A(t), B(t)$ are continuous $n \times n$ matrices. The dynamics of the system

$$\ddot{w}(t) + 2A(t)\dot{w}(t) + B(t)w(t) = 0 \tag{1}$$

exhibits nice stability properties under suitable assumptions on matrices $A(t)$ and $B(t)$. In particular, we are interested in exponential stability, that is systems for which there exist positive constants κ and M such that

$$\|\Phi(t, t_0)\| \leq Me^{-\kappa(t-t_0)} \qquad \forall t > t_0, \tag{2}$$

where Φ is the transition matrix of the system reduced to the first-order.

In order to state an important stability property we need to introduce the matrix measure ([16] chap. 2, sect. 8)

$$\mu(P) := \lim_{h \to 0^+} \frac{\|I + hP\| - 1}{h} \tag{3}$$

induced by the matrix norm $\|P\|$.

Theorem A [Sun et al, 2007]**.** *The dynamical system defined by Eq.* (1) *is exponentially stable if there exists a positive constant* m *such that*

$$l + \sqrt{l^2 + 4c} - 2m < 0 \tag{4}$$

where $l := \sup_{t \geq t_0} \max[0, 2\mu(mI - A(t))]$, $c := \sup_{t \geq t_0} \|2mA(t) - m^2 I - B(t)\|$.

Proof. See [51]. ∎

Lemma A. *Let* $\theta > 0$, *and* $B(t)$ *a continuous, diagonizable* $n \times n$ *matrix with* $B(t) = P(t)\,\mathrm{diag}(\lambda_1(t), \ldots, \lambda_n(t))P(t)^{-1}$ *such that* $\lambda_i(t) \in \mathbf{R}$ *are positive for all* $i = 1, \ldots, n$ *and all* t. *Let us consider the homogeneous differential system*

$$\ddot{\omega}(t) + \theta\dot{\omega}(t) + B(t)\omega(t) = 0. \tag{5}$$

If $\theta^2 \geq 4\lambda(1+\chi)/\chi$ *with* $\lambda := \min_i \inf_{t\geq t_0} \lambda_i(t)$ *and* $\chi := \sup_{t\geq t_0} \|P(t)\|\|P(t)^{-1}\|$, *then the dynamical system defined by Eq.* (5) *is exponentially stable.*

Proof. Once we pose $A(t) = \frac{1}{2}\theta I$ we can apply Theorem A. If we choose $m < \theta/2$ then $\mu(mI - A(t)) = \mu((m-\frac{\theta}{2})I) < 0$ and $\max[0, 2\mu((m-\frac{\theta}{2})I)] = 0$, so $l = 0$. As a consequence, for the condition (4) to be verified we need to satisfy $c < m^2$, where $c = \sup_{t\geq t_0} \|(m\theta - m^2)I - B(t)\|$. Hence, once we pose $\Lambda(t) := \mathrm{diag}(\lambda_1, \ldots, \lambda_n(t))$

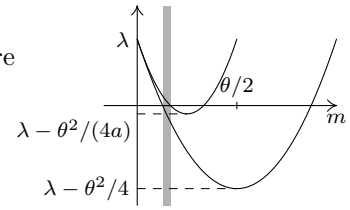
$$\begin{aligned} c &= \sup_{t\geq t_0} \|P(t)\big((m\theta - m^2)I - \Lambda(t)\big)P^{-1}(t)\| \\ &\leq \sup_{t\geq t_0} \big\|(m\theta - m^2)I - \Lambda(t)\big\| \sup_{t\geq 0} \big(\|P(t)\|\ \|P^{-1}(t)\|\big) \leq |m\theta - m^2 - \lambda|\chi, \end{aligned} \tag{6}$$

we need to satisfy

$$|m\theta - m^2 - \lambda|\chi < m^2. \tag{7}$$

Now, recalling that by assumption $\theta^2 > 4\lambda$, choose m such that $m\theta - m^2 - \lambda \geq 0$, that is to say $\theta/2 - 1/2\sqrt{\theta^2 - 4\lambda} \leq m \leq \theta/2 + 1/2\sqrt{\theta^2 - 4\lambda}$. Then Eq. (7) becomes $am^2 - \theta m + \lambda > 0$ where $1 < a := (1+\chi)/\chi$.

Since the parabolas $m^2 - \theta m + \lambda$ and $am^2 - \theta m + \lambda$ are as in the side figure, there is always at least an interval of values of m (corresponding to the grey region in the figure) that simultaneously satisfies $m\theta - m^2 - \lambda \geq 0$, $am^2 - \theta m + \lambda > 0$ and $m < \theta/2$, namely:



$$\frac{\theta - \sqrt{\theta^2 - 4\lambda}}{2} \leq m < \frac{\theta - \sqrt{\theta^2 - 4a\lambda}}{2a}. \tag{8}$$

Finally, the exponential stability of differential equation (5) follows from Theorem A. ∎

Now let $t_0 \in \mathbf{R}$, $\mathcal{E}\colon [t_0, +\infty) \to (0, +\infty)$ be a nonincreasing differentiable and integrable function such that there exists $\lim_{t\to+\infty} \mathcal{E}'(t)/\mathcal{E}(t) \in (-\kappa, 0]$,, and $\mathcal{E}(t) \geq \exp(-\kappa(t - t_0))$. A typical example is

$$\mathcal{E}(t) = \frac{\epsilon}{(\alpha + t - t_0)^p} \quad \forall t > t_0, \tag{9}$$

where $\alpha > 0$, $\epsilon > 0$, $p > 1$ are suitable parameters.

Based on classical results on the links between exponential and BIBO stability (see [12]), Theorem 1, pag. 196), the following lemma establishes a sharper property on the asymptotical relation between the input and the output of an exponentially stable system.

Lemma B. *Let* $A \in \mathcal{C}([t_0, +\infty); \mathbf{R}^{n\times n}) \cap L^\infty([t_0, +\infty); \mathbf{R}^{n\times n})$, *and suppose that the homegeneous system* $\dot{\omega}(t) = A(t)\omega(t)$ *is exponentially stable. Let* $u \in \mathcal{C}^1([t_0, +\infty); \mathbf{R}^n)$ *be such that*

$$\|u(t)\| \leq \mathcal{E}(t) \qquad \forall t > t_0. \tag{10}$$

Then there exists a positive constant K such that the solution of

$$\dot{\omega}(t) = A(t)\omega(t) + u(t) \tag{11}$$

satisfies

$$\|\omega(t)\| \leq K\mathcal{E}(t) \qquad \forall t > t_0. \tag{12}$$

Proof. For any $t > t_0$ the solution of Eq. (11) can be expressed by

$$\omega(t) = \Phi(t, t_0)\omega(t_0) + \int_{t_0}^{t} \Phi(t, \beta)u(\beta)d\beta. \tag{13}$$

Since the homogeneous system is asymptotically stable we have

$$\begin{aligned} \|\omega(t)\| &\leq \|\Phi(t, t_0)\omega(t_0)\| + \int_{t_0}^{t} \|\Phi(t, \beta)u(\beta)\| \; d\beta \\ &\leq \; M\|\omega(t_0)\|e^{-\kappa(t-t_0)} + \int_{t_0}^{t} \|\Phi(t, \beta)\| \, \|u(\beta)\| \; d\beta. \end{aligned} \tag{14}$$

Using (10), $\int_{t_0}^{t} \|\Phi(t, \beta)\| \, \|u(\beta)\| \; d\beta \leq M \; F(t)$, where we have posed $F(t) := \exp(-\kappa t) \int_{t_0}^{t} \exp(\kappa\beta)\mathcal{E}(\beta)d\beta$. An application of the De l'Hôpital theorem shows that

$$\lim_{t\to+\infty} \frac{F(t)}{\mathcal{E}(t)} = \frac{1}{\kappa + l}, \tag{15}$$

where $l := \lim_{t\to+\infty} \mathcal{E}'(t)/\mathcal{E}(t)$ and we recall that $l \in (-\kappa, 0]$. Hence from Eq. (14) there is a constant $C' > 0$ such that

$$\|\omega(t)\| \leq M\|\omega(t_0)\|e^{-\kappa(t-t_0)} + C'\mathcal{E}(t) \leq K\mathcal{E}(t), \tag{16}$$

for any $t > t_0$, where $K > 0$ is a suitable positive constant (we use here the assumption that $\mathcal{E}$ is bounded below by $\exp(-\kappa t)$). ∎

Let us consider the framework dictated by the following definition. From now on we set $t_0 = 0$.

Definition A. *An input signal $x(\cdot)$ is $\mathcal{E}$-quasiperiodic if for all $T > 0$ there exist positive constants c, $\bar{\tau}$, $\hat{\tau}_T$: $0 < \bar{\tau} < \hat{\tau}_T < T$ and a $\mathcal{C}^1$ function $\tau : [0, T - \hat{\tau}_T] \to [\bar{\tau}, \hat{\tau}_T]$ such that $\gamma(t) := t + \tau(t)$ with $\gamma' > c$ in $[\bar{\tau}, T - \hat{\tau}_T]$, $\hat{\tau}_T$ is uniformly bounded in T, and*

$$\forall t \in [0, T - \hat{\tau}_T) \text{ we have } \|x(t) - x(\gamma(t))\| \leq \mathcal{E}(t). \tag{17}$$

In the simplest case in which $\tau(t) \equiv \hat{\tau} > 0$, the above definition reduces to an extended notion of periodicity in which we are "tolerant" with respect to the match induced by the period. In general, it is interesting to pick up τ from functional spaces equipped with classical analytic properties.

Now, we will show that the assumption on quasi-periodic environments has important consequences on the learning process. In particular, we will prove that after a while the learning driven by Eq. 1.2.2–(3) has the property of convergence to coherent decisions on similar inputs; this is indeed the essence of the following theorem:

Theorem B. *Suppose:*

i. *positiveness of the spectrum of $v_{\omega\omega}$:* $\lambda := \min_{i=1,\dots,N} \inf_{t\in[0,\infty)} \lambda_i(\xi(t)) > 0$ *for all $\xi(t)$ that lie in the segment joining $(w(t), x(t))$ and $(w(\gamma(t)), x(\gamma(t)))$, where $\lambda_i(\zeta), \dots, \lambda_N(\zeta)$ are the eigenvalues of $v_{\omega_i\omega_j}(\zeta)$. Under this assumption we write $v_{\omega\omega}(\xi(t)) = P(t)\,\mathrm{diag}(\lambda_1(\xi(t)), \dots \lambda_N(\xi(t)))P(t)^{-1}$;*
ii. *$v_{\omega\xi}$ is bounded. We set $C_V := \|v_{\omega\xi}\|_\infty$;*
iii. *quasi-periodicity: $x(\cdot)$ is $\mathcal{E}$-quasiperiodic;*
iv. *global existence on $(0, +\infty)$ of the solution to Eq. 1.2.2–(3) with Cauchy initial conditions.*

Let $\theta^2 \geq 4\lambda(1+\chi)/\chi$, where $\chi := \sup_{t\in[0,T-\hat\tau_T]} \|P(t)\|\|P(t)^{-1}\|$. Then there exists $K > 0$ such that for every $t \in [0, T - \hat\tau_T]$

$$\|w(t) - w(\gamma(t))\| \leq K\mathcal{E}(t). \tag{18}$$

Proof. From Eq. 1.2.2–(3) (with $\mu \equiv 1$) at t and $\gamma(t)$ we get

$$\begin{aligned} &\ddot w(t) + \theta\dot w(t) + v_\omega(w(t), x(t)) = 0;\\ &\ddot w(\gamma(t)) + \theta\dot w(\gamma(t)) + v_\omega(w(\gamma(t)), x(\gamma(t))) = 0. \end{aligned} \tag{19}$$

Now, if we pose $\omega(t) := w(t) - w(\gamma(t))$ and

$$B(t) := \int_0^1 V_{\omega\omega}(\sigma(s,t))\,ds, \quad I(t) := \int_0^1 V_{\omega\xi}(\sigma(s,t))\,ds, \tag{20}$$

where $\sigma(s,t) := (sw(t) + (1-s)w(\gamma(t)), sx(t) + (1-s)x(\gamma(t)))$ for each $s \in [0,1]$ and $t \in [0,T]$, we have

$$v_\omega(w(t), x(t)) - v_\omega(w(\gamma(t)), x(\gamma(t))) = B(t)\omega(t) + I(t)(x(t) - x(\gamma(t))). \tag{21}$$

From Eq. (19) we get

$$\ddot\omega(t) + \theta\dot\omega(t) + B(t)\omega(t) = -I(t)(x(\gamma(t)) - x(t)). \tag{22}$$

From i. and from Lemma A, we know that the homogeneous system associated to (22) is exponentially stable. Now we can reduce Eq. (22) to the first order by setting $z_1 = \omega$ and $z_2 = \dot\omega$, so as we get

$$\dot z = -\begin{pmatrix} 0 & -\mathrm{Id} \\ B(t) & \theta\mathrm{Id} \end{pmatrix} z - \begin{pmatrix} 0 \\ I(t)\big(x(t) - x(\gamma(t))\big) \end{pmatrix} \tag{23}$$

Since $x(\cdot)$ is $\mathcal{E}$-quasiperiodic by assumption iii we have $\|I(t)\big(x(t) - x(\gamma(t))\big)\| \leq \|I(t)\|\,\|x(t) - x(\gamma(t))\| \leq C_V\mathcal{E}(t)$, for any $t \in [0, T - \hat\tau_T]$. Finally, we can apply Lemma B, and the conclusion follows. ∎

Corollary. *If v is strongly convex in ω, $v_{\omega\xi}$ is bounded, x is $\mathcal{E}$-quasiperiodic and we have global existence of Eq. 1.2.2–(3) with Cauchy initial conditions, then whenever $\theta^2 \geq 4\lambda(1+\chi)/\chi$ Then there exists $K > 0$ such that for every $t \in [0, T - \hat\tau_T]$*

$$\|w(t) - w(\gamma(t))\| \leq K\mathcal{E}(t).$$

In practice this analysis shows that the learning dynamic expressed by Eq. 1.2.2–(3) is a good candidate to model learning when the agent has a continuous dialogue with the surrounding environment. In the next chapter we will try to elaborate on this and show that indeed such dynamic can be understood also in terms of variational principles.

CHAPTER TWO

VARIATIONAL APPROACH

I have called this principle,
by which each slight variation,
if useful, is preserved,
by the term Natural Selection.

— CHARLES R. DARWIN, in *On the Origin of Species* (1859)

Nature does everything
for the sake of something

— ARISTOTELE, in *Parts of Animals* (350 BC)

Natura enim simplex est
& rerum causis superfluis non luxuriat

— Isaac Newton, in *Philosophiae Naturalis Principia Mathematica* (1687)

Caro Adso,
non occorre moltiplicare le spiegazioni e le cause
senza che se ne abbia una stretta necessità.

— Guglielmo da Baskerville in *Il nome della rosa* (1980)

Mathematically each of the three different formulations, Newton's law, the local field method and the minimum principle, gives exactly the same consequences. What do we do then? You will read in all the books that we cannot decide scientifically on one or the other. That is true. They are equivalent scientifically. It is impossible to make a decision, because there is no experimental way to distinguish between them if all the consequences are the same. But psychologically they are very different in two ways. First, philosophically you like them or do notlike them; and training is the only way to beat that disease. Second, psychologically they are different because they are completely unequivalent when you are trying to guess new laws.

— RICHARD P. FEYNMANN, in *The Character of Physical Law* (1964)

SIMPLICITY and elegance have always been an incredibly useful criterion for the development of successful theories. The underlying philosophical principle it the one that has been very clearly expressed by the English Franciscan friar William of Occam, that in his theological work *Quaestiones et decisiones in quattuor libros Sententiarum* asserts that [42]:

Pluralitas non est ponenda sine necessitate

that basically means that we do not have to choose complex explanations when we can rely on simpler ones.

One of the most fruitful ways in which this parsimony principle has been applied by scientists, is through the application of variational methods in order to describe and to develop theories in a concise and expressive way. One of the fields where the results of the application of such methods have been particularly far reaching is physics: The canonical formalism that is built starting from a functional problem is at the very heart of a great deal of the 20th century elementary particle physics. Indeed one of the reasons why physicists like so much variational approaches is the easiness with which one can directly incorporate desired symmetries and constraints. Consider, as an example, the problem of determining, in classical mechanics, the trajectory of a particle that moves freely without any force acting on it. If we assume the existence of a reference frame (inertial frame) in which space is homogeneous and isotropic and time is homogeneous, then we can (see [33]) derive the famous principle of inertia from a variational formulation without any additional hypothesis.

In Chapter 1 we saw how, at least for some tasks of interest, it is sensible to formulate a learning problem in terms of a quest for a mechanism by means of which the parameters of a learning model adapt to data. Since our ambition would be to find natural rules (laws) rather than heuristics to describe such interaction, in this chapter, we will explore a possible formulation of the problem in terms of the tools offered by the calculus of variations. We will therefore start from what we have identified as batch mode as it constitute a bridge with a mechanical system subject to drag forces without external interactions. Then we will move on to point out the main difficulties and the first results which characterize a general formulation that also takes into account interactions with a learning environment.

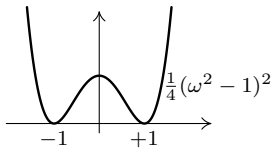
The chapter will close with a novel analysis on how neural networks can be described within a variational framework in terms of a set of honolomic constraints. This opens new possibilities to the study of deep learning and will be of crucial importance for the formulation of multilayer visual theories as we will see in Chapter 3.

2.1. A VARIATIONAL THEORY FOR BATCH MODE LEARNING

In Chapter 1 we pursue the quest for an appropriate map $t \mapsto w(t)$ that would describe the evolution of the parameters of a learning model. So basically we were looking, amongst all the possible trajectories that maps $[0,T] \to \mathbf{R}^N$ (where $[0,T]$ can be thought as the life of the agent) the "best" one with respect to a specific task expressed by a potential. We also argued in Section 1.2 that this description is precise to the extent that the potential does not depend on time, while we showed that in the case of time dependent potential we cannot formulate the problem simply in terms of a static optimization. Let us therefore investigate whether, at least in the case of time independent potential, it is possible to give a variational formulation to the problem.

The idea is the following: We want to start from a casual configuration, $w(0) = w^0$, and then find a trajectory $w(t)$ that at each time assumes the optimal value with respect to U. Clearly, stated in this terms, the problem is not well posed since, assuming that $U > 0$, this corresponds to find in some (yet to be specified) space of functions $\mathbf{X}$, a trajectory that minimizes

$$\int_0^T U(w(t))\,dt. \tag{1}$$



Without further specifications on the space $\mathbf{X}$ and on the function U there can be a lot of unwanted solutions. For example suppose that the function U has two global minima, take for example the double well function $U(\omega) := \frac{1}{4}(\omega^2 - 1)^2$, then the solution of the problem

$$\min_{w\in\mathbf{X}} \int_0^T U(w(t))\,dt, \tag{2}$$

crucially depends on the space $\mathbf{X}$. If we choose $\mathbf{X} = \{f \in L^1([0,T]) : f(0) = w^0\}$ without any other requirement in terms of regularity, whereas the problem (2) is well defined it admits infinite discontinuous solutions as for any finite partition of $[0,T]$, $t_0 = 0 < t_1 < \cdots < t_D = T$, and for any $\sigma \in \{x \in \mathbf{R}^D : x_i \in \{-1,+1\}\}$ the function

$$u(t) = \begin{cases} w^0 & \text{when } t = 0; \\ \sigma_i & \text{for } t_i < t \le t_{i+1}, \quad i = 0,\dots,D-1, \end{cases} \tag{3}$$

is a minimizer of (2) since

$$\int_0^T U(u(t))\,dt = 0. \tag{4}$$

The problem here is of course regularity; if we do not ask for higher degree of regularity the solution can change abruptly from a point to the next without paying anything. This suggests us to consider instead of (2) the following modified problem

$$\min_{w\in\mathbf{X}} \int_0^T \frac{1}{2}|\dot{w}(t)|^2 + U(w(t))\,dt, \tag{5}$$

where the presence of a penalty on the temporal derivatives of w suggests that solution (3) does not minimize (5) anymore.

Indeed assuming that the solution of this problem is regular enough, applying standard methods of calculus of variations (see Appendix A) we have that the minimum of (5) must satisfy the Euler equation

$$-\ddot{w}(t) + \nabla U(w(t)) = 0. \tag{6}$$

Assuming also that we minimize among all function which has some fixed initial value w^0 we also get that Eq. (6) must be coupled with the following boundary conditions:

$$w(0) = w^0, \qquad \dot{w}(T) = 0. \tag{7}$$

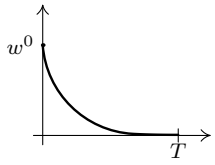
Because of the condition at $t = T$, this problem is non causal.

In order to understand this issue consider a potential for which Eq. (5) can be solved explicitly. The simplest one is the quadratic potential $U(\omega) = \frac{1}{2}\omega^2$. The solution to the following problem

$$\begin{cases} -\ddot{w}(t) + w(t) = 0; \\ w(0) = w^0; \\ \dot{w}(T) = 0, \end{cases} \tag{8}$$

is

$$w(t) = \frac{w^0}{1 + e^{2T}} e^t + \frac{w^0}{1 + e^{-2T}} e^{-t}. \tag{9}$$



As we can immediately see for all $t < T$ the divergent term e^t is actually dominated by the denominator, and therefore the solution approaches rapidly 0 which is the minimum of the potential. Moreover as $T \to +\infty$ the condition $\dot{w}(T) = 0$ becomes a condition on the integrability of the function since otherwise the integral in Eq. (5) would be divergent. If we take the limit $T \to \infty$ in Eq. (9) we can see how the divergent mode disappears leaving simply $w(t) \approx w_0 e^{-t}$.

This solution, as we have anticipated, is highly non-causal and the minimization is attained only for that initial condition on $w'(0)$, which will later yield $w'(T) = 0$. A random initialization will, in general, give to the system a divergent mode that will inevitably push the solution away from the wanted value 0.

Also the formulation of learning based on Eq. (5) therefore is not satisfactory because causality is of the utmost importance in learning. Even though biological arguments are not always completely compelling, sometimes it is useful to confront them and discuss about biological plausibility of our models. In view of this, due to a evolutionary selection criterion, one can always argue that intelligent beings actually have the correct "initialization" that will actually lead to learning. However the behaviour of differential Eq. (6) is so sensible to initial conditions that it is quite difficult to think that this is actually possible.

Causality therefore seems to be an important issue to take into account when working with minimization of integral functionals. In the next section we will devote some time to review Hamilton's principle as a formulation of classical mechanics to gain a better understanding of how Newton's laws — which are causal — relates to the corresponding variational principle of least action. Another aspect of classical mechanics on which we will comment is the minimality properties of the trajectories.

2.1.1. Variational Principle for Classical Mechanics

Variational theory of classical mechanics is based on what is known as the *Principle of least action* or *Hamilton's principle*; it is usually stated as follows [3]:

> *The solutions of the equations of Newtonian mechanic for a system described by a potential U coincides with the extremals of the functional*
>
> $$S(x) := \int_0^T \frac{1}{2}|\dot{x}|^2 - U(x) = \int_0^T L(x, \dot{x}), \tag{1}$$
>
> *where $L(z, p) := |p|^2/2 - U(z)$.*

Here it is assumed that $t \mapsto x(t)$ is the position vector of a particle in Cartesian coordinates*.

This statement is far from being precise since there are at least two important ingredients that needs to be specified:

1. The space from which $x(t)$ are drawn. This point both encompass the class of regularity and the type of boundary conditions that we assign;
2. What we mean by extremal of a functional.

It is very important to point out that in some sense 1. better define point 2. Usually in calculus of variation a (local) extremum is a (local) maximum or minimum. Global minima or maxima are very easy to define even for functions defined on sets (without further structure): $x_0 \in \mathbf{X}$ is a global minimum of $F\colon \mathbf{X} \to \overline{\mathbf{R}}$ if for every $x \in \mathbf{X}$ we have $F(x_0) \leq F(x)$. Local minima or maxima requires a notion of distances and therefore can be well defined, for example, on metric spaces (a set which have also a distance function). Indeed if we assume $x \in \mathbf{X}$, where $\mathbf{X}$ is a space without any structure, the characterization of critical points of a functional becomes quite abstract. Nevertheless one can prove that a necessary condition for x_0 to be a point of (global) minimum of F is the following: take $\delta > 0$, let $\gamma\colon (-\delta, \delta) \to \mathbf{X}$ a path in $\mathbf{X}$ such that $\gamma(0) = x_0$ and define $\varphi(s) := F(\gamma(s))$, then for any of such path we need to have

$$\varphi'(0) =: \delta F(x_0, \gamma) = 0, \tag{2}$$

whenever $\varphi'(0)$ exists.

* Also whenever it is not specified integrals are assumed to be done with Lebesgue measure. We will sometimes omit the differential symbol in the notation for the integral when we want to stress the structure of the functional as a map from a functional space to real numbers $x \mapsto S(x)$.

If we add more structure to $\mathbf{X}$, for example we assume that $(\mathbf{X}, V)$ is an affine space with vector space* V, then for all $x_0 \in \mathbf{X}$ we can consider the curve $s \mapsto \gamma(s) := x_0 + sv \in \mathbf{X}$ with $v \in V$ which is the line that passes trough x_0 with direction v. As it is explained in Appendix A the first variation of a functional in the point x_0 in the direction $v \in V$ can be defined as

$$\delta F(x_0, v) := \lim_{s \to 0} \frac{F(x_0 + sv) - F(x_0)}{s}, \tag{3}$$

which is usually known as *directional derivative* or *Gâteaux derivative*. If x_0 is an extremum of F it must be

$$\delta F(x_0, v) = 0, \quad \forall v \in V. \tag{4}$$

Nonetheless the term extremal in Hamilton's principle does not coincide with the common notion of extremals in calculus of variation (minimum or maximum), it is rather referring to what in the literature (see for example [19]) is usually called a *critical point*, that is to say a point x of $\mathbf{X}$ for which

$$\delta F(x, v) = 0, \qquad \forall v \in V. \tag{5}$$

In fact,in most cases, the trajectory of a particle does not have any minimality property with respect to the functional in (1).

As we have already remarked, the choice of $\mathbf{X}$ is particularly important because of the boundary conditions. We know that Newton's equations are second-order ODE in the trajectory variable, and because of this we also know that in mechanics the trajectory of particles is completely specified by the initial values of position and velocity. This being said it seems natural to define:

$$\mathbf{X} := \{\, x \in C^\infty([0,T]) : x(0) = x^0, \quad \dot{x}(0) = x^1 \,\}, \tag{6}$$

and

$$V := \{\, v \in C^\infty([0,T]) : v(0) = \dot{v}(0) = 0 \,\}. \tag{7}$$

We will now show that with this choice a function $x \in \mathbf{X}$ that satisfy Eq. (5) in general it does not exists.

We have (see Appendix A)

$$\begin{aligned} \delta F(x, v) &= \int_0^T L_z(x, \dot{x}) \cdot v + L_p(x, \dot{x}) \cdot \dot{v} \\ &= [L_p(x, \dot{x}) v]_0^T + \int_0^T L_z(x, \dot{x}) \cdot v - (L_p(x, \dot{x}))^{\cdot} \cdot v, \end{aligned} \tag{8}$$

where $L_z(z,p) = \nabla U(z)$, $L_p(z,p) = p$. Since stationarity condition (5) must hold for all $v \in V$ we can proceed as follows: By choosing $v \in C_c^\infty([0,T])$ we have, because of the fundamental lemma of calculus of variations,

$$\ddot{x} + \nabla U(x) = 0, \tag{9}$$

* This as we shall soon see is the case of interest here since usually $\mathbf{X}$ will be an affine subspace of an Hilbert space.

which indeed is Newton's equation for a particle that moves in the potential U, however taking $v \in V$ such that $v(T) = 1$, from (5) we also get

$$L_p(x(T), \dot{x}(T)) = 0, \quad \text{that implies} \quad \dot{x}(T) = 0. \tag{10}$$

This means that for x to be a critical point of F on the space $\mathbf{X}$ it should satisfy

$$\begin{cases} \ddot{x}(t) + \nabla U(x(t)) = 0 \quad \forall t \in (0,T); \\ x(0) = x^0; \\ \dot{x}(0) = x^1; \\ \dot{x}(T) = 0, \end{cases} \tag{11}$$

which, as can be readily understood, in general, will not have any solution. Again notice the emergence of the Neumann boundary condition at $t = T$ which ruins causality.

Hence these arguments show us that Hamilton's principle cannot be formulated in a causal way if we look for stationary points in a functional space with assigned initial position and velocity. Looking at Eq. (8) we can however understand that if we formulate the problem on the space

$$\mathbf{X} := \{\, x \in C^\infty([0,T]) : x(0) = x^0, \quad x(T) = x^T \,\}, \tag{12}$$

then a stationary point of x must satisfy

$$\begin{cases} \ddot{x}(t) + \nabla U(x(t)) = 0 \quad \forall t \in (0,T); \\ x(0) = x^0; \\ \dot{x}(T) = x^T, \end{cases} \tag{13}$$

that is indeed a well-defined boundary value problem. Thus the correct way to state Hamilton's principle would rather be the following

> *The solution of the equations of Newtonian mechanic $\bar{x}(t)$ for a system described by a potential U coincides with the stationary point (when it is unique) of the functional $S(x)$ defined over the set*
>
> $$\mathbf{X} := \{\, x \in C^\infty([0,T]) : x(0) = \bar{x}(0), \quad x(T) = \bar{x}(T) \,\}. \tag{14}$$

Rephrased in a (possibly) more appealing way this is the analogue of saying that between all possible trajectories that pass between two predetermined points, mechanics choose the one which makes the action S stationary. Again this formulation is non-causal.

Now let us now turn to the minimality properties of this functional. The simplest case is the free particle action $S(x) = \int |\dot{x}|^2/2$. In this case the stationarity conditions has as unique solution the line $x(t) = x^0 + (x^T - x^0)t/T$, which is indeed a minimum of S on $\mathbf{X}$ (see [3]). The second example is the classic harmonic oscillator that is described by the action $S(x) = \int |\dot{x}|^2/2 - |x|^2/2$; in this case stationarity conditions gives the unique point

$$x(t) = x^0 \cos t + \frac{x^T - x^0 \cos T}{\sin T} \sin t. \tag{15}$$

In order to see whether this solution is a minimum we can check whether $\delta^2 S(x, v) \geq 0$ for all $v \in V$. In this case we have

$$\delta^2 S(x, v) = \int_0^T (v \cdot L_{zz} v + 2v \cdot L_{zp} \dot{v} + \dot{v} \cdot L_{pp} \dot{v}). \tag{16}$$

Since we also have $L_{zp} = 0$, $L_{zz} = -\mathrm{Id}$, $L_{pp} = \mathrm{Id}$, then

$$\delta^2 S(x, v) = \int_0^T (|\dot{v}|^2 - |v|^2). \tag{17}$$

One can see that this quantity, with appropriate choices* of T we can always find $v \in V$ such that $\delta^2 S(x, v) < 0$; for example fix $T = 10$, then in one dimension we have

$$\delta S(x, t^2 - 10t) = -3000. \tag{18}$$

This indeed means that in this case mechanics is generally a "saddle point" of S.

Till now we saw that the variational principle in classical mechanics is non-causal, and at the same time it does not have those minimality properties that are necessary for a good theory of leaning. On the other hand the fact that energy $E(t) = \frac{1}{2}|\dot{x}(t)| + U(x(t))$ is conserved prevents the explosive behaviours that we experienced in the case of the functional in 2.1–(5). This being said we are still missing a crucial property of learning systems that we have discussed in Chapter 1: dissipation. In the remainder of this section we will try to understand if it is possible to incorporate dissipation in the action of classical mechanics.

Dissipative evolution. The main problem with the functional of mechanics described in Eq. (1) is that there is no directionality of time. Because of this, it can consistently describe a conservative system while it is rather reasonable that it cannot model dissipative systems: If we consider the motion of a conservative system (a pendulum with no friction, particles in a box that undergoes elastic scattering etc.) and shoot a film of it, then a backward reproduction of the movie will show a consistent (possible) dynamics. Instead if we consider the motion of a ball that moves in a medium and because of a drag force comes to a rest, the time-reversed dynamics would look like a ball that start accelerating by itself without an apparent cause.

So the basic idea to introduce dissipation is to insert in the action (1) a term that breaks the time reversal symmetry. Let us therefore consider a function of time $a(t)$ which is not symmetric with respect to $t = T/2$. A simple additive term will have no effects to the stationary points of the action, while if we multiply the whole Lagrangian by $a(t)$ this will affect the stationarity condition of S. So consider the following modification of the action

$$S_a(x) := \int_0^T a \left(\frac{1}{2}|\dot{x}|^2 - U(x) \right), \tag{19}$$

* This hypothesis is necessary because for T small enough solution (15) is indeed a minimum of the action; and this it may be a motivation for the name principle of least action (see for example [33] and [18])

which has as stationarity condition the following Euler equation

$$a(t)\ddot{x}(t) + \dot{a}(t)\dot{x}(t) + a(t)\nabla U(x(t)) = 0. \tag{20}$$

Assuming that $a(t) \neq 0$ for all $t \in (0, T)$ we can rewrite Eq. (20) as

$$\ddot{x}(t) + \frac{\dot{a}(t)}{a(t)}\dot{x}(t) + \nabla U(x(t)) = 0. \tag{21}$$

From this expression we can clearly see that if $\dot{a}/a > 0$ we are modelling a system with a dissipative force proportional to the velocity; the simplest case of a constant drag coefficient ν can be achieved when we choose a to be the solution of the differential equation

$$\frac{\dot{a}(t)}{a(t)} = \gamma, \tag{22}$$

which have an exponential solution $a(t) = a(0)e^{\gamma t}$. Moreover the integration constant $a(0)$ is irrelevant since it is just an overall multiplicative term. Then we are left with the choice $a(t) = e^{\gamma t}$. This construction allows us to recover the heavy-ball method we discussed in Section 1.2.1 and 1.2.2 from a variational method. However from this same variational approach we also get non-causal final condition.

The idea of weighting the standard action with an exponential to describe dissipative systems by means of a variational formulation is already present in the literature notably in [27].

2.1.2. Causal Formulation of Learning

Let us consider the functional in Eq. 2.1–(5) again. In the last section we argued that besides giving non-causal stationarity conditions, because the structure of the Lagrangian, it also does not have terms that explicitly breaks time reversal. Indeed if we consider the symmetry $t \to T - t$ (symmetry with respect to $T/2$) we have

$$\int_0^T \frac{1}{2}|\dot{w}(T-t)|^2 + U(w(T-t))\,dt = \int_T^0 \frac{1}{2}|\dot{w}(s)|^2 + U(w(s))\,(-ds). \tag{1}$$

In Section 2.1.1 we have seen that when we introduce a term that breaks this symmetry we can recover dissipative behaviour. Therefore we start to conjecture that causality can be reached with the introduction of an appropriate time dependence in the Lagrangian. So consider the following modification of the functional (similar to the one considered in Eq. 2.2.1–(19))

$$F(w) := \int_0^T \frac{a}{2}|\dot{w}|^2 + bU(w), \tag{2}$$

where $t \mapsto a(t)$ and $t \mapsto b(t)$ are two positive differentiable functions. Now consider the functional space $\mathbf{X}$ as in Eq. 2.1.1–(6) and take in this space the

variation of F. The consequent stationarity conditions are

$$\begin{cases} -a(t)\ddot{w}(t) - \dot{a}(t)\dot{w}(t) + b(t)U(w(t)) = 0; \\ a(T)\dot{w}(T) = 0, \end{cases} \tag{3}$$

where, as usual, the second condition comes from the boundary term of the integration by parts. It seems therefore that indeed we can choose the function a so to have $a(T) = 0$. If we do this, apparently the non-causal condition $a(T)\dot{w}(T) = 0$ is satisfied. However in general is never a good idea to take a weight in front of the term that enforces regularity that is vanishing, because in the region in which the weight vanishes the solution can become very large. Let us make a simple example: Consider the case of a real valued functions $w \in \mathbf{R}$, take $a(t) = b(t) = (T-t)$ and choose $U(w) = \frac{1}{2}(w)^2$. Assuming that there exists smooth stationary point we can write down the Euler equation

$$-(T-t)\ddot{w}(t) + \dot{w}(t) + (T-t)w(t) = 0. \tag{4}$$

The general solution of this equation is

$$w(t) = AI_0(T-t) + BK_0(T-t), \tag{5}$$

where I_0 and K_0 are the modified Bessel functions. This can be easily checked if we perform in Eq. (4) the change of variable $t \equiv t(s) = T - s$. If we define $z(s) := w(T-s)$ we have that z satisfies

$$s^2 z''(s) + sz'(s) - s^2 z(s) = 0, \tag{6}$$

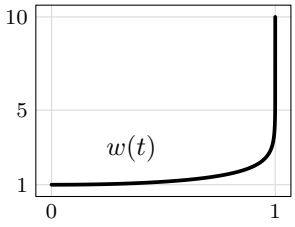
that is the defining equation of the modified Bessel functions I_0 and K_0 (see [1] pag. 374 Eq. (9.6.1)). If we now fix the integration constants A and B with Cauchy initial conditions $w(0) = w^0$ and $\dot{w}(0) = w^1$, then we find that

$$A = T(w^0 K_1(T) - w^1 K_0(T)), \quad B = T(w^0 I_1(T) + w^1 I_0(T)). \tag{7}$$

Notice that since $T > 0$ this quantities are real numbers. Therefore the solution with Cauchy initial conditions is

$$w(t) = T(w^0 K_1(T) - w^1 K_0(T))I_0(T-t) + T(w^0 I_1(T) + w^1 I_0(T))K_0(T-t). \tag{8}$$

Since the limiting form of modified Bessel functions for small arguments are known we can study the behaviour of the solution near $t = T$ which is exactly the critical point where a vanishes. In the side figure the solution is plotted for $T = 1$, $w^0 = 1$ and $w^1 = 0$. From [1] pag. 375 Eq. (9.6.7) and (9.6.8) we know that $I_0(s) \sim 1$ while $K_0(s) \sim -\log s$ as $s \to 0$, therefore we can conclude that as $t \to T$



$$w(t) \sim C_1 + C_2 \log(T-t). \tag{9}$$

This means that the solution diverges logarithmically for $t = T$ and it also means that indeed the boundary conditions are satisfied since

$$\lim_{t \to T^-} w'(t)a(t) = 0. \tag{10}$$

What about the minimality properties of this solution? Because

$$w'(t) = -AI_1(T-t) + BK_1(T-t), \tag{11}$$

and $K_1(s) \sim 1/s$ we immediately understand that the energy of this solution is divergent.

We will now show how the causality issues can be covered following an idea by Ennio de Giorgi (see [14]) that has been proved in [50], [49], [37].

De Giorgi approach. The main idea is that we can introduce a parameter ε (with the dimensions of time) which measure the "degree of causality" of the problem that we are solving. When $\varepsilon > 0$ the problems that we will define are non-causal, however as $\varepsilon \to 0$ the solutions of this problems will converge (in some sense) to the solution of a causal problem. To define this family of problems consider the functional in Eq. (2) with $a(t) \equiv a_\varepsilon(t) = \varepsilon e^{-t/\varepsilon}$ and $b(t) \equiv b_\varepsilon(t) = e^{-t/\varepsilon}$ so to have

$$F_\varepsilon(w) := \int_0^T e^{-t/\varepsilon} \left(\frac{\varepsilon}{2} |\dot{w}|^2 + U(w(t)) \right). \tag{12}$$

Let us also consider the following domain:

$$\mathbf{X} = \{ w \in H^1((0,T); \mathbf{R}^N) : w(0) = w^0 \}. \tag{13}$$

More generally (following the original conjecture by De Giorgi and its subsequent proofs) we can consider the functional

$$F_\varepsilon(w) := \int_0^T e^{-t/\varepsilon} \left(\varepsilon^2 \frac{\rho}{2} |\ddot{w}|^2 + \varepsilon \frac{\nu}{2} |\dot{w}|^2 + U(w(t)) \right) dt, \tag{14}$$

where we have introduced two positive constants to weight the regularization terms $|\ddot{w}|^2$ and $|\dot{w}|^2$; this functional is thus naturally well defined on the space

$$\mathbf{X} := \{ w \in H^2((0,T); \mathbf{R}^N) : w(0) = w^0, \quad \dot{w}(0) = w^1 \}. \tag{15}$$

The first important result that we have is that (see [37]) if $U \in C^1(\mathbf{R}^N; \mathbf{R})$ and bounded from below by using standard arguments of the direct method of calculus of variations it is possible to prove the existence of a minimizer of F_ε in $\mathbf{X}$. Let us now see what such minimizer looks like.

Let us denote w_ε a minimizer of F_ε on $\mathbf{X}$. Then surely we need to have $\delta F_\varepsilon(w_\varepsilon, v) = 0$, for all $v \in V := \{ v \in C^\infty([0,T]; \mathbf{R}^N) : v(0) = \dot{v}(0) = 0 \}$, and so

$$\delta F_\varepsilon(w_\varepsilon, v) = \int_0^T e^{-t/\varepsilon} \Big(\varepsilon^2 \rho \ddot{w}_\varepsilon(t) \cdot \ddot{v}(t) + \varepsilon \nu \dot{w}_\varepsilon(t) \cdot \dot{v}(t) + \nabla U(w_\varepsilon(t)) \cdot v(t) \Big) \, dt = 0, \tag{16}$$

which is the weak form of the Euler Lagrange equation. A boostrap argument shows also that the minimizers are also regular (in fact they are C^∞) and

therefore satisfy the strong form of the Euler equations:

$$\begin{cases} \varepsilon^2\rho w_\varepsilon^{(4)} - 2\varepsilon\rho w^{(3)} + (\rho - \nu\varepsilon)\ddot{w} + \nu\dot{w} + \nabla U(w) = 0 \quad \text{in} \quad (0,T); \\ w(0) = w^0, \qquad \rho\dot{w}(0) = \rho w^1; \\ \rho\ddot{w}(T) = 0, \qquad \rho\varepsilon w^{(3)}(T) = \nu\dot{w}(T), \end{cases} \tag{17}$$

where the two extra boundary conditions at $t = T$ arises due to boundary conditions of $v \in V$ (see Appendix A).

Let us see this in some details, as it is the crucial argument that let us conclude that the minimum in $\mathbf{X}$ has the desired regularity. Consider the following lemma

Lemma A. *Let $f \in L^2((0,T))$ and suppose that f has a second order weak derivative, i.e there exists a function $g \in L^2((0,T))$ such that $\int f\varphi'' = \int g\varphi$ for all $\varphi \in C_c^\infty((0,T))$, then $f \in H^2((0,T))$.*

Proof. It is sufficient to prove that f admits a weak derivative. Then just set $f'(t) = \int_0^t f''(t)\,dt$, then $f'(t) \in L^2((0,T))$ and for all $\varphi \in C_c^\infty((0,T))$ we have

$$\int_0^T f'\varphi' = -\int_0^T f''\varphi.$$

which concludes the proof. ∎

Equipped with this observation we can easily prove the following regularization theorem:

Theorem A. *Let w_ε be a minimum of F_ε in $\mathbf{X}$, if $U \in C^k(\mathbf{R}^N)$ then $w_\varepsilon \in C^{k+3}([0,T])$.*

Proof. From Eq. (16) we have for all $v \in C_c^\infty(0,T)$:

$$\int_0^T \varepsilon^2\rho e^{-t/\varepsilon}\dddot{w}_\varepsilon(t)\cdot\ddot{v}(t)\,dt + \int_0^T \varepsilon\nu e^{-t/\varepsilon}\ddot{w}_\varepsilon(t)\cdot\dot{v}(t)\,dt = -\int_0^T e^{-t/\varepsilon}\nabla U(w_\varepsilon(t))\cdot v(t)\,dt \tag{18}$$

which gives

$$\begin{aligned} \int_0^T \varepsilon^2\rho e^{-t/\varepsilon}\dddot{w}_\varepsilon(t)\cdot\ddot{v}(t)\,dt - \int_0^T \varepsilon\nu\left(e^{-t/\varepsilon}\ddot{w}_\varepsilon(t)\right)' v(t)\,dt \\ = -\int_0^T e^{-t/\varepsilon}\nabla U(w_\varepsilon(t))\cdot v(t)\,dt. \end{aligned} \tag{19}$$

This relation means that $e^{-t/\varepsilon}\ddot{w}_\varepsilon(t)$ has a weak second derivative and in particular

$$\frac{d^2}{dt^2}(\varepsilon^2\rho e^{-t/\varepsilon}\ddot{w}_\varepsilon(t)) = \frac{d}{dt}(\varepsilon\nu e^{-t/\varepsilon}\dot{w}_\varepsilon(t)) - e^{-t/\varepsilon}\nabla U(w_\varepsilon(t)) \quad \text{in} \quad L^2((0,T)). \tag{20}$$

Thanks to Lemma A we can therefore conclude that $\ddot{w}_\varepsilon \in H^2(0,T)$ this means that w_ε, $\dot{w}_\varepsilon$, $\ddot{w}_\varepsilon$ and $w_\varepsilon^{(3)}$ are continuous, and from Eq. (20) plus the continuity of ∇U we have that $w_\varepsilon^{(3)}$ has a continuous derivative which implies $w_\varepsilon^{(3)} \in C^1([0,T])$

or $w_\varepsilon \in C^4([0,T])$. At this point we have already obtained that (20) holds in the strong sense with "real" derivatives. Now using (20) we can express $w_\varepsilon^{(4)}$ in terms of w_ε, $\dot{w}_\varepsilon$, $\ddot{w}_\varepsilon$, so we have that $w_\varepsilon^{(4)} \in C^1([0,T])$ meaning that if $U \in C^2$ then $w_\varepsilon \in C^5([0,T])$. Repeating this same argument we have the thesis. ∎

This theorem assure us that in order to find the minimum of F_ε one can study the solutions of the system of ODE in Eq. (17).

Now the important part of De Giorgi conjecture is the fact that these minima that, as we can see from Eq. (17), are non-causal, as $\varepsilon \to 0$ converge to the solution of a causal, evolution problem that is completely determined by the initial conditions of $\mathbf{X}$. The particularly interesting fact is that the causal solution that we recover is the solution of the following Cauchy problem:

$$\begin{cases} \rho\ddot{w}(t) + \nu\dot{w}(t) + \nabla U(w(t)) = 0; \\ w(0) = w^0, \qquad \dot{w}(0) = w^1, \end{cases} \tag{21}$$

which is exactly the dynamics of the mechanical system described in the previous section (see Eq. 2.1.1–(21)). The explicit temporal dependence of the Lagrangian expressed in the term $\exp(-t/\varepsilon)$ as we already commented is necessary to have the dissipation behaviour. However in this formulation, the dissipative term $\nu\dot{w}$ can be directly linked to the term that penalized the first derivative and it emerges in the limit due to an interaction between this regularization term and the exponential factor.

The precise statement of this convergence result is summed up in the following theorem (which can be found in [37]).

Theorem B. (*Liero and Stefanelli*) *Let $\nu > 0$, and assume that w_ε minimize F_ε in $\mathbf{X}$, then $w_\varepsilon \to w$ weakly in $H^2((0,T),\mathbf{R}^N)$ (and strongly in $H^1((0,T);\mathbf{R}^N)$) if $\rho \neq 0$ and weakly in $H^1((0,T),\mathbf{R}^N)$ if $\rho = 0$, where w solves the Cauchy problem* (21).

It is important to stress the fact that he solution of Eq. (21) cannot be directly interpreted as the minimum of some functional, however it is the limit of a sequence of minima as $\varepsilon \to 0$. Therefore from our point of view it is crucial to understand how does the minimum of F_ε as $\varepsilon \to 0$ performs in terms of potential energy $U(w)$ since this is the quantity we are interested in.

Since the limiting equations are the same as the ones we considered in Chapter 1, a first answer to this question involves an energy analysis directly performed on the differential equations; because we are, for the moment, considering batch mode learning we already know that both methods (the one obtained for $\rho > 0$ and the gradient flow model obtained when $\rho = 0$) tend to dissipate energy to conquer states with lower potential.

Yet another kind of consideration seems to be relevant as well: Since the solution of the limiting Cauchy problem is approximated by minima of functionals we can safely argue that if ε is small enough, then the solution w of Eq. (21) is close to the minimum of F_ε. So the main question now is: How does the exponential factor $\exp(-t/\varepsilon)$ that, as we have seen, is fundamental for

a causal formulation of the problem does affect the behaviour of the minima? Indeed the minimization of the functional F_ε without the exponential factor would imply also a minimization of the term $\int U(w)$, whereas the introduction of an exponentially suppressing term may arise at least some concern in the interpretation of the minimum problems for ε sufficiently small.

The first thing to notice is that the minima of F_ε and $F_\varepsilon/\varepsilon$ coincides. The functional $F_\varepsilon/\varepsilon$ has, as a weight, in front of the Lagrangian the function $e^{-|t|/\varepsilon}/\varepsilon$; this weight as $\varepsilon \to 0$ is proportional to the delta measure in $t = 0$. Yet if we replace this function with another weight which still approximate the delta measure, but with compact support, the Euler-Lagrange equation will not give us any conditions on the minimum outside the support of the weight. This tells us that the weight in front the Lagrangian does not prevent the minimization process to consider the Lagrangian for all $t \in [0, T]$.

In order to get a better understanding of this limiting procedure in the remaining part of this section we will mainly focus on a the simple case $\rho = 0$, $w\colon [0, T] \to \mathbf{R}$ and $U(w) = w^2/2$, which gives first order linear problems that can be solved exactly. In this special case the Euler equation for the minimum of F_ε reads

$$\begin{cases} -\nu\varepsilon\ddot{w} + \nu\dot{w} + w = 0; \\ w(0) = w^0, \quad \dot{w}(T) = 0. \end{cases} \tag{22}$$

Let us try to compare this minimum with the one that we obtain from

$$G(w) = \int_0^T \left(\frac{\nu^2}{2} |\dot{w}|^2 + U(w(t)) \right) dt, \tag{23}$$

which solves the following problem:

$$\begin{cases} -\nu^2\ddot{w} + w = 0; \\ w(0) = w^0, \quad \dot{w}(T) = 0. \end{cases} \tag{24}$$

We want to compare the solutions of Eq. (22) in the limit $\varepsilon \to 0$ with the minimum of this functional that weights homogeneously potential U and that has the same time scale ν that is present in each F_ε. Let us call the solution of the first problem w_ε, and let ω the solution of the problem in Eq. (24). Both of then are second order differential equations with boundary conditions, however while (22) will approximate better and better the Cauchy problem (by applying Theorem B)

$$\begin{cases} \nu\dot{w} + w = 0 \\ w(0) = w^0 \end{cases} \tag{25}$$

the problem in (24) does not have a "causal approximation". Now, let us start to find the explicit solution of Eq. (22). The explicit solution for general T, ε

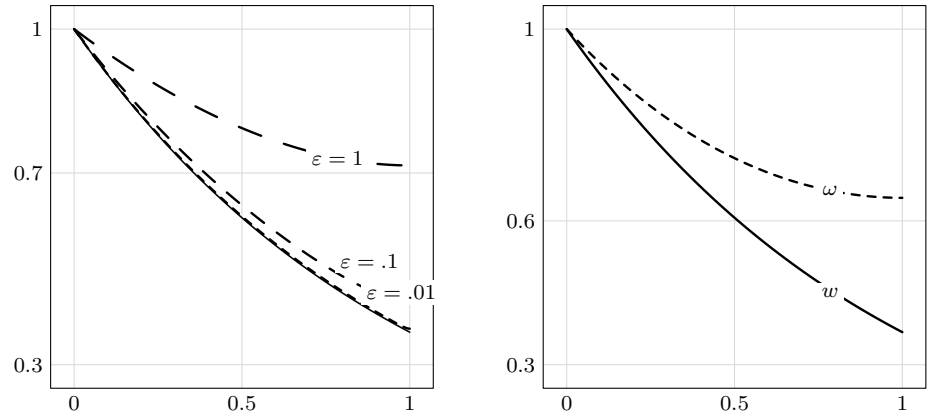


Fig. 1. On the left we plot the convergence trend of w_ε (dashed lines) the limiting value w (solid line). On the right the solution w (or w_ε for small ε) is compared against the minimum of G, ω.

and w^0, as can be easily checked by direct substitution is

$$w_\varepsilon(t) = w^0 e^{t/2\varepsilon} \frac{\sqrt{4\varepsilon+\nu}\,\mathrm{Ch}\left(\frac{\sqrt{4\varepsilon+\nu}}{2\varepsilon\sqrt{\nu}}(T-t)\right) + \sqrt{\nu}\,\mathrm{Sh}\left(\frac{\sqrt{4\varepsilon+\nu}}{2\varepsilon\sqrt{\nu}}(T-t)\right)}{\sqrt{4\varepsilon+\nu}\,\mathrm{Ch}\left(\frac{\sqrt{4\varepsilon+\nu}}{2\varepsilon\sqrt{\nu}}T\right) + \sqrt{\nu}\,\mathrm{Sh}\left(\frac{\sqrt{4\varepsilon+\nu}}{2\varepsilon\sqrt{\nu}}T\right)}. \quad (26)$$

Instead the solution $\omega(t)$ is simply given by

$$\omega(t) = w^0 \frac{1}{\mathrm{Ch}(T/\sqrt{\nu})}\,\mathrm{Ch}\left(\frac{T-t}{\sqrt{\nu}}\right) \quad (27).$$

The behaviour of this solutions is displayed in Fig. 1.; Fig. 1–(a) shows the convergence trend of $w_\varepsilon \to w$, instead Fig. 1–(b) compares the behaviour of a minimum of the functional G (dashed line) against the limiting solution w obtained solving Eq. (25)

As it can be seen the qualitative behaviour of the solutions is pretty similar even though ω is the solution of a second order equation (with boundary conditions), while w comes from a first order ODE (with Cauchy conditions) and in both cases the solutions lower the potential energy.

For equations with one main condition
(Those linear), you have the permission
To take your solutions,
With firm resolution,
And add them in superposition.

— DAVID MORIN, *Introduction to Classical Mechanics* (2007)

Of course the same analysis can be carried out explicitly in the case $\rho > 0$ with quadratic potential; in this case the solution is a superposition of ex-

ponential modes $w_\varepsilon(t) = \sum_{i=1}^{4} c_i(\varepsilon) e^{\lambda_i(\varepsilon) t}$, where $\lambda_i(\varepsilon)$ are the roots of the characteristic polynomial

$$\varepsilon^2 \rho z^4 - 2\varepsilon\rho z^3 + (\rho - \varepsilon\nu) z^2 + \nu z + 1. \tag{28}$$

As it is well known an explicit algebraic solution for this roots can be found and, in particular in this case the roots turns out to be

$$\begin{aligned} \lambda_1 &= \frac{1 - \sqrt{1 + (2\varepsilon/\rho)\big(\nu - \sqrt{\nu^2 - 4\rho}\big)}}{2\varepsilon}, \quad \lambda_2 = \frac{1 - \sqrt{1 + (2\varepsilon/\rho)\big(\nu + \sqrt{\nu^2 - 4\rho}\big)}}{2\varepsilon}, \\ \lambda_3 &= \frac{1 + \sqrt{1 + (2\varepsilon/\rho)\big(\nu - \sqrt{\nu^2 - 4\rho}\big)}}{2\varepsilon}, \quad \lambda_4 = \frac{1 + \sqrt{1 + (2\varepsilon/\rho)\big(\nu + \sqrt{\nu^2 - 4\rho}\big)}}{2\varepsilon}. \end{aligned} \tag{29}$$

Eventually we were therefore able to find a variational scheme which produces causal/evolution laws, that converges to low potential states with some temporal regularity on the variations of the parameters w (we worked out the solution only for quadratic potentials but general convergence properties are classical results and can be found for example in [26]) which is indeed what we hoped for.

We also argue that the limiting solution that comes out of this method cannot be interpreted as the minimum of some functional; indeed we have that $F_\varepsilon / \to U(w^0)$ whose minimum clearly is not the solution of (25). In the next section we will discuss whether it is possible to achive similar causality properties but with a different kind of explicit temporal dependence; possibly a temporal dependence that weights more evenly the potential on the whole temporal interval.

2.1.3. Why the Exponential?

In the previous section we considered a sequence of F_ε based on the exponential weight $\exp(-t/\varepsilon)$. We asked ourselves whether the causal reduction that we discussed in Section 2.1.2 heavily depends on the exponential form or it the same behaviour could be reproduced for a broader family of weights?

So let replace in the definition of F_ε the term $e^{-t/\varepsilon}$ with some arbitrary positive function $t \mapsto \varpi_\varepsilon(t) > 0$. If we assume that such a weight is regular (for example C^∞) then all the regularity results that we gave in the last section will still hold true and the Euler equations for a fixed ε will turn out to be

$$\varepsilon^2 \rho \varpi_\varepsilon w^{(4)} + 2\varepsilon^2 \rho \dot{\varpi}_\varepsilon w^{(3)} + (\varepsilon^2 \rho \ddot{\varpi}_\varepsilon - \varepsilon\nu\varpi_\varepsilon)\ddot{w} - \varepsilon\nu\dot{\varpi}_\varepsilon \dot{w} + \varpi_\varepsilon \nabla U = 0, \tag{1}$$

with the two boundary conditions at $t = T$

$$\begin{cases} \rho\varepsilon^2 \varpi_\varepsilon(T)\ddot{w}(T) = 0 \\ \rho\varepsilon^2 w^{(3)}(T) + \rho\varepsilon^2 \dfrac{\dot{\varpi}_\varepsilon(T)}{\varpi_\varepsilon(T)} \ddot{w}(T) - \nu\varepsilon\dot{w}(T) = 0 \end{cases} \quad \Rightarrow \quad \begin{cases} \ddot{w}(T) = 0 \\ \rho\varepsilon w^{(3)}(T) = \nu\dot{w}(T) \end{cases} \tag{2}$$

Now, since we assumed ϖ_ε to be positive (remember that in Section 2.1.2 we discussed how it is not a very good idea to let the weight become zero) we can

rewrite Eq. (1) in the following way:

$$\varepsilon^2\rho w^{(4)} + 2\varepsilon^2\rho\frac{\dot{\varpi}_\varepsilon}{\varpi_\varepsilon}w^{(3)} + \Big(\varepsilon^2\rho\frac{\ddot{\varpi}_\varepsilon}{\varpi_\varepsilon} - \varepsilon\nu\Big)\ddot{w} - \varepsilon\nu\frac{\dot{\varpi}_\varepsilon}{\varpi_\varepsilon}\dot{w} + \nabla U = 0. \tag{3}$$

Therefore the limiting dynamic will crucially depends on the form of two functions: $\dot{\varpi}_\varepsilon/\varpi_\varepsilon$ and $\ddot{\varpi}_\varepsilon/\varpi_\varepsilon$. More specifically it will depend on the form of these two functions as $\varepsilon \to 0$. First of all notice that the term $\dot{\varpi}_\varepsilon/\varpi_\varepsilon$ appears both multiplied by ε and ε^2 therefore if we assume that $\dot{\varpi}_\varepsilon/\varpi_\varepsilon \sim f(t)/\varepsilon^2$, then there would surely be a problem with the term $\varepsilon\dot{\varpi}_\varepsilon/\varpi_\varepsilon$ because it would be divergent. This is in general what will happen if we take $\dot{\varpi}_\varepsilon/\varpi_\varepsilon \sim f(t)/\varepsilon^n$ with $n > 1$. Moreover If we choose $n < 1$ both the terms $\varepsilon\dot{\varpi}_\varepsilon/\varpi_\varepsilon$ and $\varepsilon^2\dot{\varpi}_\varepsilon/\varpi_\varepsilon$ would vanish. So the only remaining possibility is to choose

$$\dot{\varpi}_\varepsilon(t)/\varpi_\varepsilon(t) \sim f(t)/\varepsilon, \quad \text{as } \varepsilon \to 0. \tag{4}$$

Looking at the differential equation we also want to have $f(t) < 0$ for all $t \in [0, T]$, otherwise, as we have already commented, the dynamic can be unstable. Then, once we choose f (assuming that this choice should be smooth) we can find the appropriate weight by solving the differential equation

$$\dot{\varpi}_\varepsilon(t) = \frac{f(t)}{\varepsilon}\varpi_\varepsilon(t). \tag{5}$$

This equation can be easily solved by choosing $\varpi_\varepsilon(t) = \exp(g(t))$, where $g(t)$ solves $\dot{g} = f(t)/\varepsilon$. Hence if we let $F(t) = F(0) + \int_0^t f(s)\,ds$ we finally have that

$$\varpi_\varepsilon(t) = \exp\Big(\frac{F(0)}{\varepsilon} + \frac{1}{\varepsilon}\int_0^t f(s)\,ds\Big), \tag{6}$$

where $F(0)$ has to do with the initial condition on the weight. Now since we required f to be negative, the second term in the exponential is negative, the first term though depends on the sign of $F(0)$. If we allow $F(0)$ to be strictly positive as $\varepsilon \to 0^+$ then the weight will blow up, which is an unwanted behaviour. Then we assume that $F(0) \leq 0$. Under this hypothesis however for any smooth choice of f we will have the same exponential dumping behaviour that we wanted to avoid with the $e^{-t/\varepsilon}$.

This arguments all together are a strong indication that with a smooth weight we either get a trivial limiting equation or (if we choose the weight more wisely) we end up in a situation that has the same exponential damping properties that we wanted to avoid in the first place.

Discontinuous weight. The next question therefore is to see whether something change if we consider weights that allows discontinuities. The basic idea being the following: Instead of using one exponential term $e^{-t/\varepsilon}/\varepsilon$ that in the limit $\varepsilon \to 0$ sample the value of the Lagrangian (and hence of the potential) at $t = 0$, if we allow discontinuities we can consider a whole "train" of this terms. As $\varepsilon \to 0$ if the space between two of these peaks (see Fig. 2) also goes to 0 then at the end we can transform the integral that defines the functional that we are considering into a uniform integral over the interval $[0, T]$ of the potential.

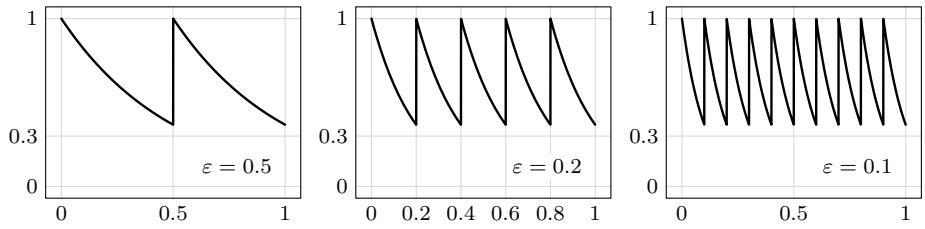


Fig. 2. The function $\varpi_\varepsilon(t)$ written in Eq. (7) for three different values on ε when $T = 1$.

So, without further ado let us define the following discontinuous weight function*:

$$\varpi_\varepsilon(t) = e^{-t/\varepsilon} \sum_{k=0}^{T/\varepsilon-1} e^k \chi_{[t_k,t_{k+1})}(t), \quad t_k = \varepsilon k, \tag{7}$$

where χ_I is the characteristic function of the interval I.

Once we defined this weight then we can consider the family of functionals

$$\begin{aligned} J_\varepsilon(w) &:= \int_0^T \varpi_\varepsilon\Big(\varepsilon^2\frac{\rho}{2}|\ddot{w}|^2 + \varepsilon\frac{\nu}{2}|\dot{w}|^2 + U(w)\Big) \\ &= \sum_{k=0}^{T/\varepsilon-1} e^k \int_{t_k}^{t_{k+1}} e^{-t/\varepsilon}\Big(\varepsilon^2\frac{\rho}{2}|\ddot{w}|^2 + \varepsilon\frac{\nu}{2}|\dot{w}|^2 + U(w)\Big)\,dt. \end{aligned} \tag{8}$$

First of all we need to know if this modification has compromised the result on the existence of the minima on the set $\mathbf{X}$ (see Eq. 2.1.2–(15)).

The proof of the existence of the minima (under the same hypothesis used in [37]) can be established using the direct method of calculus of variations.

Theorem C. *For every $\varepsilon > 0$ the problem $\min\{J_\varepsilon(w) : w \in \mathbf{X}\}$ has a solution.*

Proof. The change of the weight surely does not change the results on lower semicontinuity, thus we just need to check the compactness of sublevels of J_ε with respect to the convergence $w_n \to q$ strongly in $H^1((0,T);\mathbf{R}^n)$ and $\ddot{w}_n \rightharpoonup \ddot{q}$ weakly in $L^2((0,T);\mathbf{R}^n)$.

First of all let us define

$$J_\varepsilon(w;(a,b)) := \int_a^b e^{-t/\varepsilon}\Big(\varepsilon^2\frac{\rho}{2}|\ddot{w}|^2 + \varepsilon\frac{\nu}{2}|\dot{w}|^2 + U(w)\Big)\,dt, \tag{9}$$

then we have that

$$J_\varepsilon(w;(a,b)) \geq e^{-b/\varepsilon}\varepsilon^2\rho\int_a^b |\ddot{w}|^2\,dt \tag{10}$$

* From now on we will always assume that ε divides T, that is to say we choose only the ε such that $T/\varepsilon \in \mathbf{N}$.

Now it is clear that* if $J_\varepsilon(w) \le C$ then $J_\varepsilon(w;(a,b)) \le C$, therefore

$$e^{-b/\varepsilon}\varepsilon^2\rho\int_a^b |\ddot{w}|^2\,dt \le J_\varepsilon(w;(a,b)) \le C, \tag{11}$$

Hence

$$\begin{aligned} C \ge J_\varepsilon(w) = \sum_{k=0}^{T/\varepsilon-1} e^k J_\varepsilon(w;(t_k,t_{k+1})) &\ge \sum_{k=0}^{T/\varepsilon-1} e^k e^{-(k+1)}\varepsilon^2\rho\int_{t_k}^{t_{k+1}} |\ddot{w}|^2\,dt \\ &= \frac{\varepsilon^2\rho}{e}\int_0^T |\ddot{w}|^2\,dt, \end{aligned} \tag{12}$$

which gives us the estimate

$$\int_0^T |\ddot{w}|^2\,dt \le \frac{Ce}{\rho\varepsilon^2}. \tag{13}$$

Moreover in virtue of the following inequality (proved in Lemma 2.3 of [49])

$$\int_0^T e^{-t}|h(t)|^2\,dt \le 2|h(0)|^2 + 4\int_0^T e^{-t}|h'(t)|^2\,dt \tag{14}$$

with $h(t) = \dot{w}(t)$ and with $\dot{w}(0) = w_1$ we have also

$$\int_0^T |\dot{w}|^2\,dt \le C. \tag{15}$$

Now we can use the fundamental theorem of calculus to estimate $w(t)$ for $t \in [0,T]$ as follows:

$$w(t) = w_0 + \int_0^T \dot{w}\,dt \le w_0 + \sqrt{T}\left(\int_0^T |\dot{w}|^2\,dt\right). \tag{16}$$

Once we observe this we can follow a standard coercivity argument; first of all we notice that any sequence $\langle w_n\rangle$ is bounded in $H^2([0,T];\mathbf{R}^N)$ we have that $\|w_n\|_{H^2} < C$ for all $n \in \mathbf{N}$. Then because of Theorem 3.16 in [10] it follows that $\langle w_n\rangle$ has a subsequence weakly converging in $H^2((0,T);\mathbf{R}^N)$. Moreover since H^2 compactly embeds in H^1, then there is a subsequence that converges strongly in H^1. This means that indeed the sublevels of J_ε are compact with respect to the notion of convergence specified earlier. ∎

So this proves that the problem at least is well posed. We should also comment on the regularity of the solutions; since we are in H^2 the solution must be at least continuous and differentiable; however globally we cannot say more this because of the discontinuities that the weight ϖ_ε. Nonetheless we can repeat the argument of Theorem B choosing v to be compactly supported in each (t_k,t_{k+1}); in doing so we establish the regularity of the solution on each of the intervals where the weight is regular. So the only points where we need to

* Throughout the proof the value that we assign to the constant C can change. Of course it is intended that it is a constant that does not depend on the index of the sequence n.

understand carefully how the solution behaves are the t_k points. In order to do this, we need, first of all, to understand how to treat the variation of an integral functional with a time-dependent lagrangian that have explicit discontinuities in time.

To evaluate the first variation of J_ε it is better to take a slightly general point of view and consider:

$$J_\varepsilon(w) = \int_0^T L_\varepsilon(t, w(t), \dot{w}(t), \ddot{w}(t))\, dt, \tag{17}$$

where $L_\varepsilon(t,\cdot,\cdot,\cdot)$ is discontinuous at times $t = t_k$, for $k = 1, 2, \ldots, T/\varepsilon$. For any $v \in V = \{C^\infty((0,T); \mathbf{R}^N) \mid v(0) = \dot{v}(0) = 0\}$ we have that

$$\begin{aligned}\delta J_\varepsilon(w, v) = \int_0^T & (L_x(t, w(t), \dot{w}(t), \ddot{w}(t))v(t) + L_p(t, w(t), \dot{w}(t), \ddot{w}(t)))\dot{v}(t) \\ & + L_a(t, w(t), \dot{w}(t), \ddot{w}(t))\ddot{v}\, dt;\end{aligned} \tag{18}$$

since the Lagrangian is discontinuous at $t = t_k$ for all $k = 1, 2, \ldots, T/\varepsilon$ we cannot directly integrate by parts, we can however split the integral into the sum of integrals each performed on a continuous Lagrangian and then integrate by parts:

$$\begin{aligned}\delta J_\varepsilon(w, v) &= \sum_{k=1}^{T/\varepsilon} \int_{t_{k-1}}^{t_k} \left(L_x v + L_p \dot{v} + L_a \ddot{v}\right) dt \\ &= \sum_{k=1}^{T/\varepsilon} \int_{t_{k-1}}^{t_k} \left(L_x - \frac{d}{dt}L_p + \frac{d^2}{dt^2}L_a\right) v\, dt + \sum_{k=1}^{T/\varepsilon} \left[L_a \dot{v} + \left(L_p - \frac{d}{dt}L_a\right)v\right]_{t_{k-1}^+}^{t_k^-}.\end{aligned} \tag{19}$$

We know that the minimum is regular in each interval (t_{k-1}, t_k), therefore by imposing $\delta J_\varepsilon(w, v) = 0$ with the variation $v \in C_c^\infty(t_{k-1}, t_k; \mathbf{R}^N)$ for $k = 1, 2, \ldots, T/\varepsilon$ one obtains the Euler equations

$$L_x(t, w(t), \dot{w}(t), \ddot{w}(t)) - \frac{d}{dt}L_p(t, w(t), \dot{w}(t), \ddot{w}(t)) + \frac{d^2}{dt^2}L_a(t, w(t), \dot{w}(t), \ddot{w}(t)) = 0, \tag{20}$$

that holds $\forall t \in \bigcup_{k=1}^{T/\varepsilon}(t_{k-1}, t_k)$. Once we know this for all the other possible variations which are not compactly supported in $\bigcup_{k=1}^{T/\varepsilon}(t_{k-1}, t_k)$ we have that

$$\begin{aligned}\delta J_\varepsilon(w, v) &= \sum_{k=1}^{T/\varepsilon} \left[L_a \dot{v} + \left(L_p - \frac{d}{dt}L_a\right)v\right]_{t_{k-1}^+}^{t_k^-} \\ &= L_a\big|_{t=T}\dot{v}(T) + \left(L_p - \frac{d}{dt}L_a\right)\Big|_{t=T} v(T) + \sum_{k=2}^{T/\varepsilon-1} \left(L_a\big|_{t=t_k^-} - L_a\big|_{t=t_k^+}\right)\dot{v}(t_k) \\ &\quad + \sum_{k=2}^{T/\varepsilon-1} \left[\left(L_p - \frac{d}{dt}L_a\right)\Big|_{t=t_k^-} - \left(L_p - \frac{d}{dt}L_a\right)\Big|_{t=t_k^+}\right] v(t_k) = 0,\end{aligned} \tag{21}$$

where we have used the continuity of v and $\dot{v}$. Appropriate choices of the variation makes us conclude that

$$\begin{aligned}
&L_a\big|_{t=T} = \Big(L_p - \frac{d}{dt}L_a\Big)\Big|_{t=T} = 0;\\
&L_a\big|_{t=t_k^-} = L_a\big|_{t=t_k^+} \quad k = 1, 2, \dots, T/\varepsilon - 1;\\
&\Big(L_p - \frac{d}{dt}L_a\Big)\Big|_{t=t_k^-} = \Big(L_p - \frac{d}{dt}L_a\Big)\Big|_{t=t_k^+} \quad k = 1, 2, \dots, T/\varepsilon - 1.
\end{aligned} \tag{22}$$

Now coming back to our original problem, we need to evaluate this conditions on the specific Lagrangian

$$L(\tau, x, p, a) = \varpi_\varepsilon(\tau)\left(\frac{\rho\varepsilon^2}{2}|a|^2 + \frac{\nu\varepsilon}{2}|p|^2 + U(x)\right). \tag{23}$$

Equation (20) of course reduces exactly to Eq. (1) and it holds for each $t \neq t_k$; Notice that for each $t \in [0, T]$ with $t \neq t_k$ we have that

$$\frac{\dot{\varpi}_\varepsilon}{\varpi_\varepsilon} = -\frac{1}{\varepsilon}, \quad \text{and} \quad \frac{\ddot{\varpi}_\varepsilon}{\varpi_\varepsilon} = \frac{1}{\varepsilon^2}. \tag{24}$$

Hence, outside the discontinuities we have that for each fixed ε the equations that holds are the Euler equations that one would obtain with $e^{-t/\varepsilon}$, i.e. the one described in Eq. 2.1.2–(17). Instead the boundary and transmission condition in Eq. (22) becomes

$$\begin{cases}
\rho\varepsilon^2\varpi_\varepsilon(T)\ddot{w}(T) = 0;\\
\nu\varepsilon\varpi_\varepsilon(T)\dot{w}(T) - \rho\varepsilon^2\dot{\varpi}_\varepsilon(T)\ddot{w}(T) - \rho\varepsilon^2\varpi_\varepsilon(T)w^{(3)}(T) = 0;\\
\rho\varepsilon^2\big(\varpi_\varepsilon(t_k^-)\ddot{w}(t_k^-) - \varpi_\varepsilon(t_k^+)\ddot{w}(t_k^+)\big) = 0 \quad k = 1, \dots T/\varepsilon - 1;\\
\nu\varepsilon(\varpi_\varepsilon(t_k^-)\dot{w}(t_k^-) - \varpi_\varepsilon(t_k^+)\dot{w}(t_k^+)) - \rho\varepsilon^2(\dot{\varpi}_\varepsilon(t_k^-)\ddot{w}(t_k^-)\\
\qquad + \varpi_\varepsilon(t_k^-)w^{(3)}(t_k^-) - \dot{\varpi}_\varepsilon(t_k^+)\ddot{w}(t_k^+) - \varpi_\varepsilon(t_k^+)w^{(3)}(t_k^+)) = 0 \quad k = 1, \dots, T/\varepsilon - 1.
\end{cases} \tag{25}$$

The first two conditions are the boundary conditions at $t = T$ and in the case $\varpi_\varepsilon(t) = e^{-t/\varepsilon}$ reduces exactly to the boundary conditions that defines the problem in Eq. 2.1.2–(17). The second two equations instead are the transmission conditions that prescribe the behaviours of the solution around the critical points t_k. Together with this conditions we have to remember that the this solution is in H^2 and therefore it must be continuous with continuous derivative everywhere. In particular this also means that the following conditions should hold:

$$w(t_k^+) = w(t_k^-) \quad \text{and} \quad \dot{w}(t_k^+) = \dot{w}(t_k^-). \tag{26}$$

So far we have formulated the minimum problem for J_ε, we have shown the existence of a solution of such a problem and we have determined the condition (differential, boundary and transmission conditions) that such solution must satisfy. The next step to follow the De Giorigi approach would be that of studying this solutions, $w_\varepsilon(t)$ as $\varepsilon \to 0$. It is important to bear in mind that this procedure

(letting $\varepsilon \to 0$) is linked to causality, as we have already remarked many times for any $\varepsilon > 0$ the solution w_ε is not causal.

Based on the analysis that we have done so far we can put out two different conjectures on the limiting behaviour of w_ε:

Conjecture A. *Let w_ε be a minimizer of J_ε in the class* $\mathbf{X}$*, then there exists* $\lim_{\varepsilon\to 0} w_\varepsilon(t) = w(t)$ *where $w(t)$ satisfy 2.1.2–*(21).

Conjecture B. *Let w_ε be a minimizer of J_ε in the class* $\mathbf{X}$*, then there exists* $\lim_{\varepsilon\to 0} w_\varepsilon(t) = w(t)$ *where $w(t)$ satisfy* $\nabla U(w(t)) = 0$.

Clearly these two conjectures are mutually exclusive (except for the very special case $\rho = \nu \equiv 0$). In the remaining of this section we will go on and investigate this two conjectures by trying to understand what happens in for quadratic potentials $U(w) = w^2/2$. In order to simplify boundary conditions we will also choose $\nu = 0$ and $N = 1$ (but of course the following argument can be carried on with $\nu \neq 0$ and for finite values of N the only important assumption here is on the form of the potential).

The Euler equation (1) and the boundary/transmission conditions (25) in this case simply becomes

$$\begin{cases} \varepsilon^2\rho w^{(4)}(t) - 2\varepsilon\rho w^{(3)}(t) + \rho\ddot{w}(t) + w(t) = 0, \qquad \forall t \in \bigcup_{k=1}^{T/\varepsilon}\big((k-1)\varepsilon, k\varepsilon\big); \\ w(0) = w^0, \qquad \dot{w}(0) = w^1; \\ \ddot{w}(T) = w^{(3)}(T) = 0; \\ \ddot{w}(k\varepsilon^+) = \dfrac{1}{e}\ddot{w}(k\varepsilon^-), \qquad k = 1, \dots T/\varepsilon - 1; \\ w^{(3)}(k\varepsilon^+) = \dfrac{1}{e}w^{(3)}(k\varepsilon^-), \qquad k = 1, \dots T/\varepsilon - 1. \end{cases} \tag{27}$$

Let us call w_ε the solution of this problem. Notice that we already know that this solution in the k-th interval $((k-1)\varepsilon, k\varepsilon)$ coincide with the function

$$w_k(t) = c_{4k-3}e^{\lambda_1(t-(k-1)\varepsilon)} + c_{4k-2}e^{\lambda_2(t-(k-1)\varepsilon)} + c_{4k-1}e^{\lambda_3(t-(k-1)\varepsilon)} + c_{4k}e^{\lambda_4(t-(k-1)\varepsilon)}, \tag{28}$$

for suitable constants $c_1, c_2, \dots, c_{4T/\varepsilon}$ and λ_i are the roots specified in Eq 2.1.2–(29). This being said the solution of the problem Eq. (27) is equivalent to the solution of the following linear system in the $4T/\varepsilon$ variables c_i

$$\begin{cases} w_1(0) = w^0; \\ \dot{w}_1(0) = w^1; \\ \ddot{w}_{T/\varepsilon}(T) = 0; \\ w^{(3)}_{T/\varepsilon}(T) = 0; \\ \ddot{w}_{k+1}(k\varepsilon) = \dfrac{1}{e}\ddot{w}_k(k\varepsilon), \qquad k = 1, \dots, T/\varepsilon - 1; \\ w^{(3)}_{k+1}(k\varepsilon) = \dfrac{1}{e}w^{(3)}_k(k\varepsilon), \qquad k = 1, \dots, T/\varepsilon - 1, \end{cases} \tag{29}$$

where the last two equations represent the transmission conditions at each discontinuous points $t = t_k$.

Another, more viable approach for solving this problem is to use the final conditions on $\ddot{w}$ and $w^{(3)}$ to determine c_1, c_2, c_3 and c_4 imposing only Cauchy conditions on the first interval and then use the continuity of w and $\dot{w}$ and the transmission conditions on $\ddot{w}$ and $w^{(3)}$ to determine the initial condition for the next interval. In the first interval the coefficients will be given by

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \end{pmatrix}^{-1} \begin{pmatrix} w^0 \\ w^1 \\ \ddot{w}_1(0) \\ w_1^{(3)}(0) \end{pmatrix}. \tag{30}$$

Then the solution and its derivative at $t = \varepsilon$ can be found simply by multiplying the appropriate exponential factors by the relative constants

$$\begin{pmatrix} w_1(\varepsilon) \\ \dot{w}_1(\varepsilon) \\ \ddot{w}_1(\varepsilon) \\ w_1^{(3)}(\varepsilon) \end{pmatrix} = \begin{pmatrix} e^{\lambda_1\varepsilon} & e^{\lambda_2\varepsilon} & e^{\lambda_3\varepsilon} & e^{\lambda_4\varepsilon} \\ \lambda_1 e^{\lambda_1\varepsilon} & \lambda_2 e^{\lambda_2\varepsilon} & \lambda_3 e^{\lambda_3\varepsilon} & \lambda_4 e^{\lambda_4\varepsilon} \\ \lambda_1^2 e^{\lambda_1\varepsilon} & \lambda_2^2 e^{\lambda_2\varepsilon} & \lambda_3^2 e^{\lambda_3\varepsilon} & \lambda_4^2 e^{\lambda_4\varepsilon} \\ \lambda_1^3 e^{\lambda_1\varepsilon} & \lambda_2^3 e^{\lambda_2\varepsilon} & \lambda_3^3 e^{\lambda_3\varepsilon} & \lambda_4^3 e^{\lambda_4\varepsilon} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \end{pmatrix}^{-1} \begin{pmatrix} w^0 \\ w^1 \\ \ddot{w}_1(0) \\ w_1^{(3)}(0) \end{pmatrix} \tag{31}$$

Define $\alpha(\ddot{w}_1(0), w_1^{(3)}(0)) := (w^0, w^1, \ddot{w}_1(0), w_1^{(3)}(0))'$, let V be the Vandermonde matrix, and $(E)_{ij} = \lambda_j^{i-1} \exp(\lambda_j \varepsilon)$ and $R = \mathrm{diag}(1, 1, 1/e, 1/e)$. Then applying the same arguments that led to equations (30) and (31) we have that

$$\begin{pmatrix} w_2(\varepsilon) \\ \dot{w}_2(\varepsilon) \\ \ddot{w}_2(\varepsilon) \\ w_2^{(3)}(\varepsilon) \end{pmatrix} = REV^{-1}\alpha(\ddot{w}_1(0), w_1^{(3)}(0)), \tag{32}$$

and again that

$$\begin{pmatrix} w_3(2\varepsilon) \\ \dot{w}_3(2\varepsilon) \\ \ddot{w}_3(2\varepsilon) \\ w_3^{(3)}(2\varepsilon) \end{pmatrix} = (REV^{-1})^2\alpha(\ddot{w}_1(0), w_1^{(3)}(0)), \tag{33}$$

up to

$$\begin{pmatrix} w_{T/\varepsilon}((T/\varepsilon - 1)\varepsilon) \\ \dot{w}_{T/\varepsilon}((T/\varepsilon - 1)\varepsilon) \\ \ddot{w}_{T/\varepsilon}((T/\varepsilon - 1)\varepsilon) \\ w_{T/\varepsilon}^{(3)}((T/\varepsilon - 1)\varepsilon) \end{pmatrix} = (REV^{-1})^{T/\varepsilon - 1}\alpha(\ddot{w}_1(0), w_1^{(3)}(0)). \tag{34}$$

Now we can also define $\omega(w_{T/\varepsilon}(T), \dot{w}_{T/\varepsilon}(T)) = (w_{T/\varepsilon}(T), \dot{w}_{T/\varepsilon}(T), 0, 0)'$ thus having

$$\omega(w_{T/\varepsilon}(T), \dot{w}_{T/\varepsilon}(T)) = EV^{-1}(TEV^{-1})^{T/\varepsilon - 1}\alpha(\ddot{w}_1(0), w_1^{(3)}(0)), \tag{35}$$

which is a 4×4 linear system with unknowns $\ddot{w}_1(0)$, $w_1^{(3)}(0)$, $w_{T/\varepsilon}(T)$ and $\dot{w}_{T/\varepsilon}(T)$. By closer inspection we realize that the system decouples; let us define

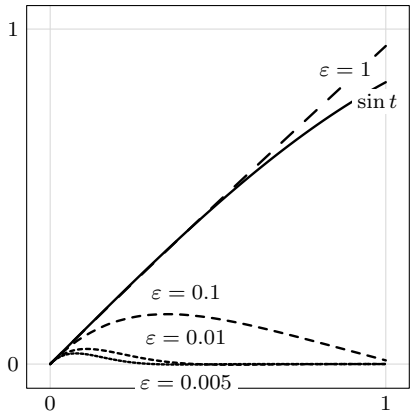


Fig. 3. Solution of Eq. (29) for $T = 1$ and $\varepsilon = 1, 0.1, 0.01, 0.005$ for $w^0 = 0$ and $w^1 = 1$.

$M := EV^{-1}(REV^{-1})^{T/\varepsilon-1}$ with entries m_{ij} and let us look at the last two lines of equation (35):

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix} \begin{pmatrix} w^0 \\ w^1 \\ \ddot{w}_1(0) \\ w_1^{(3)}(0) \end{pmatrix}, \tag{36}$$

which is equivalent to the 2×2 system

$$\begin{pmatrix} m_{33} & m_{34} \\ m_{43} & m_{44} \end{pmatrix} \begin{pmatrix} \ddot{w}_1(0) \\ w_1^{(3)}(0) \end{pmatrix} = \begin{pmatrix} -m_{31}w^0 - m_{32}w^1 \\ -m_{41}w^0 - m_{42}w^1 \end{pmatrix}. \tag{37}$$

Once we know the values of $\ddot{w}_1(0)$ and $w_1^{(3)}(0)$ we can find the coefficients $c^k := (c_{4k-3}, c_{4k-2}, c_{4k-1}, c_{4k})'$ relative to the k-th interval as follows:

$$c^k = V^{-1}(REV^{-1})^{k-1}\alpha(\ddot{w}_1(0), w_1^{(3)}(0)). \tag{38}$$

Notice that the matrix M propagates the final conditions back through each interval.

A simulation of the solution obtained in this way for the problem defined on the interval $[0, 1]$ is shown in Fig. 3. As it can be seen the solution as $\varepsilon \to 0$ becomes closer and closer to the null solution $w \equiv 0$. In this simulation the initial conditions where chosen to be $w^0 = 0$ and $w^1 = 1$, therefore the solution of Eq. 2.1.2–(21) (with $\nu = 0$) would be the $w(t) = \sin t$. This result of course gives us strong indications that Conjecture A is false while Conjecture B might be correct. Such result is not good for our purposes since the equation $\nabla U = 0$ does not give us "update rules" to modify the parameters of a learning model. We will now briefly explore another possibility that it might deserve some further investigation.

ε-independence. The above discussion show how formulating the problem subdividing the temporal interval in small segments of length ε, and then letting $\varepsilon \to 0$ probably yields solutions that converge to a value of w for which $\nabla U = 0$.

If we inspect closely the linear example discussed in the previous section, it seem clear that this behaviour is practically enforced by the transmission conditions that at each transition point t_k reduce the derivatives of the solution of a factor $1/e$.

This seems to suggest that we may avoid this behaviour (and in some sense encourage causality) by defining in each ε-interval an independent minimum problem. More specifically, suppose that we define

$$J_\varepsilon^k(w) := \int_{k\varepsilon}^{(k+1)\varepsilon} \varpi_\varepsilon \Big(\varepsilon^2 \frac{\rho}{2} |\ddot{w}|^2 + \varepsilon \frac{\nu}{2} |\dot{w}|^2 + U(w) \Big), \tag{39}$$

and consider for each $k = 0, 1, \ldots, T/\varepsilon - 1$ the problem

$$P_\varepsilon^k : \min_{w \in \mathbf{X}_\varepsilon^k} J_\varepsilon^k(w), \tag{40}$$

with

$$\mathbf{X}_\varepsilon^k := \begin{cases} \{w \in H^2(0,\varepsilon) : w(0) = w^0, \dot{w}(0) = w^1\} & \text{if } k = 0 \\ \{w \in H^2(t_k, t_{k+1}) : w(k\varepsilon) = w_\varepsilon^k(k\varepsilon), \dot{w}(k\varepsilon) = \dot{w}_\varepsilon^k(k\varepsilon)\} & \text{if } k > 0 \end{cases} \tag{41}$$

where we have set w_ε^k to be the solution of the problem P_ε^{k-1}. Notice that the sequence of problems P_ε^k is well defined since the first problem P_ε^0 has solution that defined the values of $w_\varepsilon^1(\varepsilon)$ and $\dot{w}_\varepsilon^1(\varepsilon)$, that in turns define the problem P_ε^1 and so on.

Another important comment is the fact that the function on $[0, T]$ that is constructed by concatenating the solutions of the problem P_ε^k that is to say the function

$$\bar{w}_\varepsilon(t) := \begin{cases} w^0 & \text{if } t = 0; \\ w_\varepsilon^1(t) & \text{if } 0 < t \le \varepsilon; \\ \vdots & \\ w_\varepsilon^{T/\varepsilon}(t) & \text{if } T - \varepsilon < t \le T, \end{cases} \tag{42}$$

is a competitor of the global minimum of J_ε since we have that $\bar{w}_\varepsilon \in \mathbf{X}$. This indeed follows from the following lemma:

Lemma B. *if $f \in H^2(a,b) \cap H^2(b,c)$, $f(c^+) = f(c^-)$, and $f'(c^+) = f'(c^-)$ then $f \in H^2(a,b)$.*

Proof. All we have to do is to show that there exists $g \in L^2(a,b)$ such that $\forall \varphi \in C_c^2(a,b)$, we have $\int_a^b f\varphi'' = \int_a^b g\varphi$. Indeed

$$\int_a^b f\varphi'' = \int_a^c f\varphi'' + \int_c^b f\varphi'' = -\int_a^c f'\varphi' - \int_c^b f'\varphi' = +\int_a^c f''\varphi + \int_c^b f''\varphi, \tag{43}$$

so it is sufficient to choose

$$g = \begin{cases} f'' & \text{in } (a,c) \\ f'' & \text{in } (c,b) \end{cases} \in L^2(a,b) \quad \blacksquare \tag{44}$$

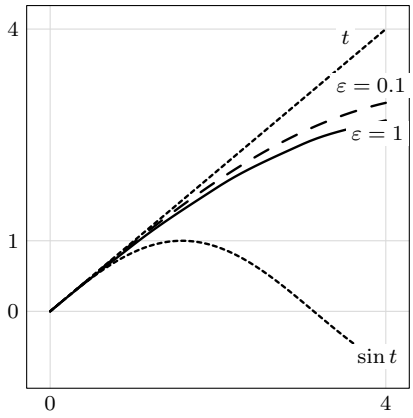


Fig. 4. Solutions in the interval $[0, 4]$ of the problems P_ε^k with quadratic potential and $\nu = 0$ for $\varepsilon = 1, 0.1,$; the reference solutions $\sin t$ and t are displayed in dashed lines.

Exactly as we did for the J_ε minima also in this case we would like to understand what happens when $\varepsilon \to 0$ since indeed each of the problem P_ε^k for each fixed ε are non-causal problems. And exactly as we did for the global minimum of J_ε we can study explicitly the problem in the case $\nu = 0$, $N = 1$ and $U = w^2/2$. This time however the coefficients of the general form of the solution described in Eq. (28) are not determined taking into account what happens too far into the future, instead are determined in each interval by the value of the solution and its derivative in the previous interval and by the boundary conditions at the end of the current ε-interval. The constant c_k that determine the solution w_ε^k should satisfy the system

$$\begin{cases} w_1(0) = w^0; \\ \dot{w}_1(0) = w^1; \\ w_k((k-1)\varepsilon) = w_{k-1}(k\varepsilon), \quad k = 2, \dots, N; \\ \dot{w}_k((k-1)\varepsilon) = \dot{w}_{k-1}(k\varepsilon), \quad k = 2, \dots, N; \\ \ddot{w}_k(k\varepsilon) = 0, \quad k = 1, \dots, N; \\ w_k^{(3)}(k\varepsilon) = 0, \quad k = 1, \dots, N, \end{cases} \tag{45}$$

where w_k is defined in Eq. (28) in terms of the coefficients c_k. Figure 4 shows the result of a simulation for $T = 4$, $w(0) = w^0$ and $\dot{w}(0) = w^1$ of the solution of the sequence of problems P_ε^k for various values of ε. From this simulation it seems clear that as $\varepsilon \to 0$ the solution converges to the solution of the Cauchy problem

$$\begin{cases} \ddot{w}(t) = 0; \\ w(0) = w^0, \\ \dot{w}(0) = w^1, \end{cases}$$

which of course is a causal problem but indeed it is not a good solution for our purposes .

2.2. A THEORY FOR TIME DEPENDENT POTENTIAL

In the last section we analyzed variational techniques for a class of fairly special potentials; the one that in the first chapter we have called time-independent potential. We argued that this kind of potential correspond in machine learning to a batch mode. Now we turn into the exploration of potentials which involve explicit dependence on time (see also [54]). Of course, since time-independent potentials are just a particular case of potentials with explicit temporal dependence the analysis carried out in the previous section, especially the discussion on causality, is relevant for time dependent potentials. This suggests us that we should start right away to consider the approach conjectured by De Giorgi and already discussed in the previous section.

So the first thing that we are going to do is to analyze what happens when we introduce an explicit temporal dependence in the functional 2.1.2–(14).

2.2.1. Causality

We will now consider the functional

$$\bar{F}_\varepsilon(w) := \int_0^T e^{-t/\varepsilon} \left(\varepsilon^2 \frac{\rho}{2} |\ddot{w}|^2 + \varepsilon \frac{\nu}{2} |\dot{w}|^2 + U(w(t), t) \right) dt, \tag{1}$$

which is the same as the one in Eq. 2.1.2–(14) except for the fact that we now allow the Lagrangian to have an extra explicit temporal dependence thought the potential U.

The temporal dependence of course does not change the result on the existence of a minimum on $\mathbf{X}$ as long as we maintain the hypothesis on boundedness from below, the regularity assumptions; the same comment goes for Theorem A. This means that the Euler equations for the minimizer are

$$\begin{cases} \varepsilon^2 \rho w_\varepsilon^{(4)}(t) - 2\varepsilon\rho w^{(3)}(t) + (\rho - \nu\varepsilon)\ddot{w}(t) + \nu\dot{w}(t) + \nabla U(w,t) = 0 \quad \text{in} \quad (0,T); \\ w(0) = w^0, \qquad \rho\dot{w}(0) = \rho w^1; \\ \rho\ddot{w}(T) = 0, \qquad \rho\varepsilon w^{(3)}(T) = \nu\dot{w}(T), \end{cases} \tag{2}$$

Moreover if we assume the following additional growth conditions on U, that is to say for all δ positive there exists two positive integrable functions $c_\delta(t)$ and $\kappa_\delta(t)$ such that for every $z \in \mathbf{R}^N$ and for all $t \in [0,T]$ we have

$$|\nabla U(z,t)| \le \delta(U(z,t) + |z|^2) + c_\delta(t), \quad |\partial_t U(z,t)| \le \delta(U(z,t) + |z|^2) + \kappa_\delta(t), \tag{3}$$

then we can prove the following theorem:

Theorem D. *The solution of the problem in Eq.* (2) *converges (weakly in* $H^1((0,T), \mathbf{R}^N)$ *to the solution of*

$$\begin{cases} \rho\ddot{w}(t) + \nu\dot{w}(t) + \nabla U(w(t), t) = 0; \\ w(0) = w^0, \qquad \dot{w}(0) = w^1, \end{cases} \tag{4}$$

Proof. The proof of this theorem follows the spirit of Theorem 4.2 of [37]. We will start with an uniform (in ε) estimate of $\|\dot{w}_\varepsilon\|_{L^2}^2$ and then we will use this

estimate in weak form of the Euler equation to show the convergence of w_ε to the solution of (4). We will prove the theorem in the case $\rho > 0$ and $\nu = 0$.
Uniform Estimate. Start form the differential equation in (2) and scalar multiply it by $(w'_\varepsilon - w^1)$:

$$\varepsilon^2 \rho w^{(4)}_\varepsilon \cdot (w'_\varepsilon - w^1) - 2\varepsilon\rho w^{(3)} \cdot (w'_\varepsilon - w^1) + \rho \ddot{w} \cdot (w'_\varepsilon - w^1) + \nabla U \cdot (w'_\varepsilon - w^1) = 0,$$

then integrate this equation on the interval $(0, t)$, and using the boundary conditions (2) integrate by parts to obtain

$$\begin{aligned}
&\varepsilon^2 \rho w^{(3)}_\varepsilon(t) \cdot (w'_\varepsilon - w^1) - \frac{\varepsilon^2\rho}{2}|\ddot{w}_\varepsilon(t)|^2 + \frac{\varepsilon^2\rho}{2}|\ddot{w}_\varepsilon(0)|^2 \\
-2\varepsilon\rho w^{(3)}_\varepsilon(t) &\cdot (\dot{w}_\varepsilon(t) - w^1) + 2\varepsilon\rho \int_0^t |\dot{w}_\varepsilon(s)|^2 \, ds + \frac{\rho}{2}|\dot{w}_\varepsilon(t) - w^1|^2 \\
+U(w_\varepsilon(t), t) - U(w^0, 0) &- \int_0^t \nabla U(w_\varepsilon(s), s) \cdot w^1 \, ds - \int_0^t \partial_t U(w_\varepsilon(s), s) \, ds.
\end{aligned}$$

Now let us integrate this equality again in the interval $(0, T)$, therefore obtaining

$$\begin{aligned}
&\left(2\varepsilon - \frac{3}{2}\varepsilon^2\right) \int_0^T \rho|\ddot{w}_\varepsilon(s)| \, ds + \frac{\varepsilon^2(1+T)}{2}\rho|\ddot{w}_\varepsilon(0)| + \left(\frac{1}{2} - \varepsilon\right)\rho|\dot{w}_\varepsilon(T) - w^1|^2 \\
+2\varepsilon\rho &\int_0^T \int_0^\tau \ddot{w}_\varepsilon(s) \, ds d\tau + \frac{\rho}{2}\int_0^T |\dot{w}_\varepsilon(s) - w^1|^2 \, ds + U(w_\varepsilon(T), T) \\
+ \int_0^T U(w_\varepsilon(s), s) \, ds &= \int_0^T \nabla U(w_\varepsilon(s), s) \cdot w^1 + \int_0^T \int_0^\tau \nabla U(w_\varepsilon(s), s) \cdot w^1 \, ds d\tau \\
+(1+T)U(w^0, 0) &+ \int_0^T \int_0^\tau \partial_t U(w_\varepsilon(s), s) \, ds d\tau.
\end{aligned}$$

Now we can take all the positive (for ε small enough) terms to the right hand side to obtain

$$\begin{aligned}
\frac{\rho}{2}\int_0^T |\dot{w}_\varepsilon - w^1|^2 \, dt + \int_0^T U(w_\varepsilon(t), t) \, dt \leq &(1+T)U(w^0, 0) \\
&+ (1+T)|w^1| \int_0^T |\nabla U(w_\varepsilon(t), t)| \, dt \\
&+ T \int_0^T |\partial_t U(w_\varepsilon(t), t)| \, dt.
\end{aligned}$$

Now using Eq. (3) we can choose δ to further reduce this inequality down to

$$\frac{\rho}{2}\int_0^T |\dot{w}_\varepsilon - w^1|^2 \, dt + \int_0^T U(w_\varepsilon(t), t) \, dt \leq c(T) + C(T) \int_0^T |w_\varepsilon(t)| \, dt,$$

where $c(T)$ and $C(T)$ are constant with respect to the parameter ε. Using Peter-Paul inequality we have that $|\dot{w}_\varepsilon - w^1|^2 \geq (1 - \eta')|\dot{w}_\varepsilon|^2 + (1 - 1/\eta')|w^1|$ for all $\eta' > 0$. Similarly since $w_\varepsilon \in H^2$, we can write $w_\varepsilon(t) = w^0 + \int_0^t \dot{w}_\varepsilon$

and using Peter-Paul and Cauchy-Schwartz we also end up with the estimate $|w_\varepsilon - w^0| \geq (1-\eta)|w_\varepsilon| + (1-1/\eta)|w^0|$ for all $\eta > 0$, which implies

$$\int_0^T |w_\varepsilon(t)|^2\,dt \leq T\frac{1/\eta - 1}{1-\eta}|w^0|^2 + \frac{T^2}{1-\eta}\int_0^T |\dot{w}_\varepsilon(t)|^2\,dt.$$

Finally this gives us the wanted uniform bound on $\|\dot{w}_\varepsilon\|_{L^2}$.
Convergence. Once we have this uniform bound we can complete the proof by arguing along the vary same line of the proof of Theorem 3.1 of [37] to obtain the thesis. ∎

In the same spirit of what we have done in Section 2.1.2 we will show an application of this result to a simple time dependent potential. We will start with what we can call a one dimensional linear auto-encoder defined by the potential

$$U(w,t) := \frac{1}{2}\sin^2 t(w-1)^2. \tag{5}$$

What we are thinking here is to have a single linear neuron which mimics a sinusoidal input. As we did in section 2.1.2 we will focus on the case $\rho = 0$ and $\nu > 0$; in this way we will have a limiting first order differential equation and second order problems for $\varepsilon > 0$ which are much simpler to handle rather than fourth order equations. The approximating problems therefore are

$$\begin{cases} -\nu\varepsilon\ddot{w}(t) + \nu\dot{w}(t) + \sin^2 t(w(t)-1) = 0; \\ w(0) = w^0; \\ \dot{w}(T) = 0, \end{cases} \tag{6}$$

while the limiting problem reduces to

$$\begin{cases} \nu\dot{w}(t) + \sin^2 t(w(t)-1) = 0; \\ w(0) = w^0. \end{cases} \tag{7}$$

The solution of Eq. (6) cannot be easily expressed in terms of elementary special function and since for our purposes it is not essential to find the exact solution we will not spend trying to find the explicit solution to this equation and we will settle for numerical solutions. Equation (7) instead is a first order linear equation that can be easily solved explicitly and it gives

$$w(t) = 1 + (w^0 - 1)\exp\Big(-2t + \frac{\sin(2t)}{4\nu}\Big). \tag{8}$$

This solution is particularly interesting as it is easily interpretable. We can immediately see that the solution converges (as it is expected) to the constant solution $w(t) = 1$, in this way the value of $w(t)\sin t$ (the output of our linear neuron) will be equal to the input $\sin t$. Moreover we can see that if we were lucky enough to start form the initial condition $w^0 = 1$, then the solution would not have been changed at all. In Fig. 5–(a) we can see how the approximating problem converges to the solution of Eq. (7).

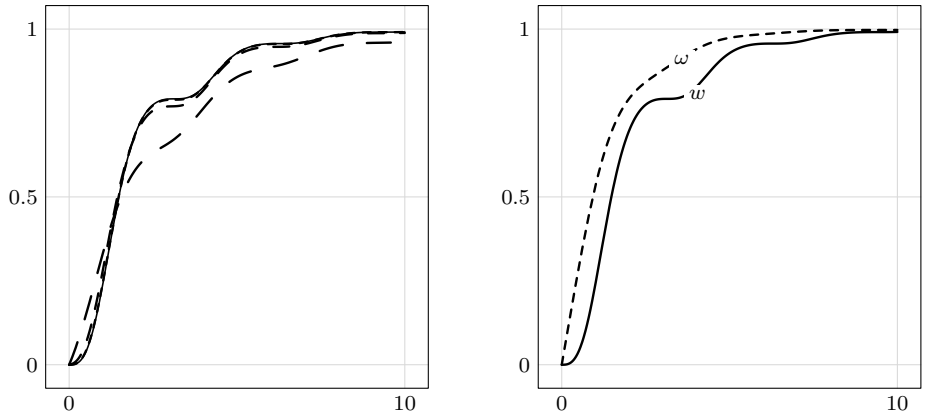


Fig. 5. On the left we plot the convergence trend of w_ε (dashed lines) the limiting value w (solid line). On the right the solution w (or w_ε for small ε) is compared against the minimum of G, ω.

Now let compare this solution with the non causal solution that comes from the minimization of

$$\bar{G}(w) = \int_0^T \left(\frac{\nu^2}{2}|w|^2 + U(w(t), t) \right) dt. \tag{9}$$

The Euler equations for this functional are

$$\begin{cases} -\nu^2 \ddot{w} + \sin^2 t(w(t) - 1) = 0; \\ w(0) = w^0, \quad \dot{w}(T) = 0. \end{cases} \tag{10}$$

A numerical solution of this problem is plotted in Fig. 5–(b) when $T = 10$ $w^0 = 0$, $\nu = 1$ and it is compared against the solution of Eq. (8) with the same parameters. We refer to the solution of (6) as w_ε, to the solution of (7) as w and we let ω be the solution of (10). Again we can see that the behaviour of the non-causal minimum and the causal one are very similar even though one solution comes from a second order differential equation (ω) and the other is the solution of a first order differential equation (w).

We will now show how, when an explicit time dependence is present, the temporal scale defined by the explicit temporal dependence of the potential furnish a natural way to segment the flux of information. Therefore this additional information from the environment opens the doors of new possible causal variational interpretations.

2.2.2. Input Segmentation

In section 2.1.2 we saw that the formulation of a causal theory with vanishing ε intervals would not give meaningful results for learning while it would have given us a nicer interpretation in terms of the quality of the solution (measured

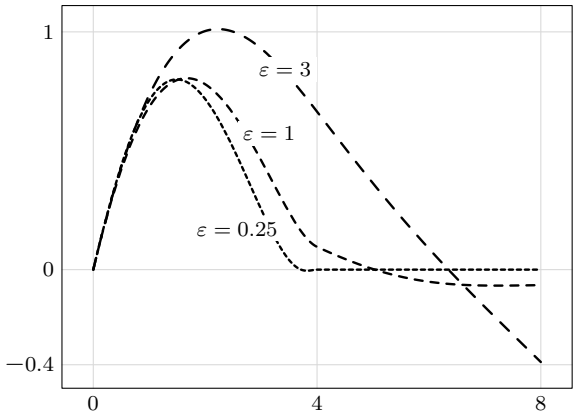


Fig. 6. Solution for $T = 8$, $\nu = 0$ of

$$\arg\min \int \varpi_\varepsilon \Big(\frac{\dot{w}^2}{2} + \frac{w^2}{2} \Big),$$

when $\varepsilon = 3, 1, 1/4$ and w_ε as in Eq. (1) with $K = 1$, $t_1 = T/2 = 4$.

through the potential). We will now show that when we can partition the temporal interval $[0, T]$ into a number of *finite* sub-intervals, then we can furnish a different interpretation of the limiting causal problem. The main observation is the following: The term ε has of course the dimensions of a time, therefore when we one consider the limit $\varepsilon \to 0$ we are actually considering the regime when $\varepsilon \ll T$. If we now have another temporal scale τ then we can replicate the problem on $[0, T]$ in each of the intervals of size τ, causality will then be achieved when $\varepsilon \ll \tau$. Of course the interpretation that we will find will crucially depend on the partition that we perform in the interval $[0, T]$ so this partition must be chosen carefully. In the batch mode problem that we discussed in the previous section the only temporal constant that we can use to perform this partition is given by a suitable combination of the parameters that we use (so on the values of ν and ρ). It is much more interesting (and indeed is the reason why we are presenting this ideas in this section) when the temporal scale is chosen based on the explicit temporal dynamic of the potential.

The idea that we would like to follow is the same that we pursued in Section 2.1.3. Consider a finite partition of size K of $[0, T]$: $t_0 = 0 < t_1 < t_2 < \cdots < t_K$. Then we can define the following weight

$$\varpi_\varepsilon(t) = e^{-t/\varepsilon} \sum_{k=0}^{K} e^{t_k/\varepsilon} \chi_{[t_k, t_{k+1})}(t), \tag{1}$$

which is the analogous of the weight defined in Eq. 2.1.3–(7) but where now t_k are fixed and they do not depend on ε. First of all we will give a counterexample that shows that even with a finite partition the limiting solution of global (on $[0, T]$) minima is not a good solution for learning.

Indeed consider the simplest nontrivial partition $0 < T/2 < T$ and the potential $w^2/2$. In Fig. 6 it is shown the behaviour if the minima w_ε of the functional $\int_0^T \varpi_\varepsilon(|\dot{w}|/2 + |w|^2/2)$ on the usual set $\mathbf{X}$. As it can be seen the solutions has a peculiar trend as $\varepsilon \to 0$; the solution converges to $w \equiv 0$ right after the discontinuity of the weight ϖ_ε.

However we can state a meaningful result under the assumption of what we can refer to as *sequential optimization* that is basically the procedure introduced in Section 2.1.3 and which assume to solve a (finite) sequence of problems one after the other and the solution to the k-th problem determines the initial conditions of the $k+1$-th problem. In Section 2.1.3 we defined each problem P_ε^k (see Eq. (40)) on a temporal domain of vanishing size; this time we will define each problem in the exactly same way:

$$P_\varepsilon^k : \min_{w \in \mathbf{X}_\varepsilon^k} J_\varepsilon^k(w), \quad k = 1, \dots, K \tag{2}$$

except that this time the functional J_ε^k is defined on a fixed temporal domain

$$J_\varepsilon^k(w) := \int_{t_{k-1}}^{t_k} \varpi_\varepsilon(t) \Big(\varepsilon^2 \frac{\rho}{2} |\ddot{w}(t)|^2 + \varepsilon \frac{\nu}{2} |\dot{w}(t)|^2 + U(w(t), t) \Big) \, dt, \tag{3}$$

and the set $\mathbf{X}_\varepsilon^k \equiv \mathbf{X}^k$ and it is defined as

$$\mathbf{X}^k := \begin{cases} \{w \in H^2(0, t_1) : w(0) = w^0, \dot{w}(0) = w^1\} & \text{if } k = 1 \\ \{w \in H^2(t_{k-1}, t_k) : w(t_{k-1}) = w_\varepsilon^k(t_{k-1}), \dot{w}(t_{k-1}) = \dot{w}_\varepsilon^k(t_{k-1})\} & \text{if } k > 1 \end{cases} \tag{4}$$

where w_ε^k is the solution of the problem P_ε^k. Also let

$$\bar{w}_\varepsilon(t) := \begin{cases} w^0 & \text{if } t = 0; \\ w_\varepsilon^k(t) & \text{if } t_{k-1} < t \le t_k. \end{cases} \tag{4}$$

Then we conjecture that

$$\bar{w}_\varepsilon \to w,$$

where w solves 2.2.1–(4). Indeed we can formally prove this in the following special case

Theorem E. *Let $\rho = 0$, $\nu > 0$ and choose $U(w,t) = w^2/2$, then $w_\varepsilon \to w$ pointwise in $[0, T]$.*

Proof. It is sufficient to prove the property for a partition of two intervals $0 < \bar{t} < T$; we also set $\nu = 1$. The characteristic polynomial associated with the ODE in (4) is $-\varepsilon\lambda^2 + \lambda + 1$, so that the roots are $\lambda_{1,2} = (1 \pm \sqrt{1 + 4\varepsilon})/2\varepsilon$. The solution in the first interval in terms of the boundary conditions is

$$w_\varepsilon^1(t) = w^0 \frac{\lambda_2 e^{\lambda_2 \bar{t}}}{\lambda_2 e^{\lambda_2 \bar{t}} - \lambda_1 e^{\lambda_1 \bar{t}}} e^{\lambda_1 t} + w^0 \frac{\lambda_1 e^{\lambda_1 \bar{t}}}{\lambda_1 e^{\lambda_1 \bar{t}} - \lambda_2 e^{\lambda_2 \bar{t}}} e^{\lambda_2 t}.$$

On the second interval the solution is

$$\begin{aligned} w_\varepsilon^2(t) =& w^0 \frac{e^{(\lambda_1+\lambda_2)\bar{t}}(\lambda_1 - \lambda_2)}{\lambda_1 e^{\lambda_1 \bar{t}} - \lambda_2 e^{\lambda_2 \bar{t}}} \frac{\lambda_2 e^{\lambda_2 \bar{t}}}{\lambda_2 e^{\lambda_2 \bar{t}} - \lambda_1 e^{\lambda_1 \bar{t}}} e^{\lambda_1 t} \\ &+ w^0 \frac{e^{(\lambda_1+\lambda_2)\bar{t}}(\lambda_1 - \lambda_2)}{\lambda_1 e^{\lambda_1 \bar{t}} - \lambda_2 e^{\lambda_2 \bar{t}}} \frac{\lambda_1 e^{\lambda_1 \bar{t}}}{\lambda_1 e^{\lambda_1 \bar{t}} - \lambda_2 e^{\lambda_2 \bar{t}}} e^{\lambda_2 t}. \end{aligned}$$

Now, by direct calculation we see that $w_\varepsilon^1(t) \to w^0 e^{-t}$ and similarly $w_\varepsilon^2(t) \to w^0 e^{-t}$ which is indeed the solution to (4). ∎

So we found that, not only 2.2.1–(4) can be interpreted as the limit of minima of $\bar{F}_\varepsilon$ on $\mathbf{X}$, but also as the solution of the "sequential" problems (2).

So let us now discuss how we can define a good partition of the interval $[0, T]$ based on the temporal evolution of the environment.

Choosing a meaningful partition. A natural way to define a partition of $[0, T]$ looking at the properties of the input signal $x(t)$ that characterize the temporal dependence of the potential is to segment according to the variations of x.

In Chapter 1 we stressed the importance of using smooth signals $x(t)$. This assumption is now fundamental in the definition of an input-based partition. Indeed if the signal is at least differentiable we for every chosen $\delta > 0$ we can always find a partition $t_0 = 0 < t_1 < \cdots < t_K = T$ of $[0, T]$ such that for every $t_k < t_a, t_b < t_{k+1}$, $k = 0, \ldots, K-1$ we have

$$\|x(t_a) - x(t_b)\| \leq \delta. \tag{5}$$

The idea behind this definition is that the partition can be done in such a way that within each interval the value of the input is almost constant. Of course in order to be able to do this we also need to have a smooth dependence of the potential on the input signal.

Once we choose such partition, the function in Eq. (4) can be interpreted as the parameter trajectory that in each of the intervals of the partition is moved according to the input present in that interval conditionally to the value of the parameters developed up to the previous interval. Indeed thanks to the weight ϖ_ε for ε small enough the solution $\bar{w}_\varepsilon$ solves minimization problems that fully "see" the data since the value of the potential evaluated at t_k, $k = 0, \ldots K$ is computed according to the value of the signal at those same temporal instants.

Furthermore the criterion that we have described so far can be further enriched to encompass the case when the input signal $x(t)$ has a finite number of discontinuities by simply inserting the times at which the discontinuities occurs into the temporal partition without changing the position of the problems P^k_ε. Sometimes abrupt (or very quick) changes indeed have the function of segmentation; as we will see in the next chapter focus of attention for visual scenes have exactly this kind of behaviour. So it is very important to have a theory that can handle such phenomena.

Another interesting comment is the fact that the solution of the problems P^k_ε (that as we have already remarked in Section 2.1.3, Lemma B as a whole belongs to $\mathbf{X}$) becomes the actual global minimum if on the boundaries of the partition the transmission conditions in Eq. 2.1.3–(25) are satisfied. As we have argued in the last section in general this does not happen, however one may imagine that with a suitable temporal dependence of the potential there could be some relaxation properties through which the derivative of the solution vanish. In this case indeed the transmission condition would be satisfied and the independence condition that we have postulated in the formulation of the problems P^k_ε would be actually transformed into a "reset" property of the derivative of the solution at

special temporal instants. This reset mechanism is exactly what we will enforce in the first formulation of our theory for visual feature extraction presented in Section 3.2.

2.2.3. Comments on the Solution

What we have done till now shows that the requirement of causality and having enforced locality by assuming that the parameters of the model w appears in the Lagrangian evaluated at a specific point in time (and not for example as a sum or integral on larger temporal domains) give rise ultimately to the differential laws 2.2.1–(4).

However we didn't have done any assumption on the information content of the signal $x(t)$. Without assumptions on the way in which the environment acts on the system there is no guarantee of good learning behaviour nor of any kind of convergence of the solutions of Eq. 2.2.1–(4). We have already seen in Section 1.2.3 that this kind of results can be achieved assuming some kind of repetitions of important pattern; for example we proved that under the assumption of quasi-periodicity the dynamic of Eq. 2.2.1–(4) becomes consistent on similar patters.

When we set $\rho = 0$ and $\nu > 0$, as we have already commented in Chapter 1, Eq. 2.2.1–(4) can be actually be interpreted as the continuum version of a SGD update. In this case it is experimentally and theoretically well known that the learning rate should be small (see [41]) and therefore ν (which has the dimensions of time) should be large compared to the temporal scale at which the example present themselves.

A last comment here that has to do with the behaviour of solutions of Eq 2.2.1–(4) is also very relevant: For the purpose of learning the speed of the signal can be adjusted by performing a smoothing on data. This is particularly interesting because it allows to avoid, or at least reduce at the beginning of learning a temporal dynamics of the gradient that are too quick therefore enhancing the quality of the numerical solution.

2.3. CONSTRAINED VARIATIONAL PROBLEMS

One of the many advantages of using calculus of variations to describe a theory, other than its compactness and the way in which symmetries are easily handled, is the fact that it allows the study of constrained problems. Indeed there is an infinite dimensional extension of the Lagrange multipliers method that guarantees (under some suitable hypothesis) the existence of the multipliers associated to a class of constraints.

In this section we will explore a method to explicitly compute these multipliers for a sufficiently big class of functionals. We will start with a somewhat general setting and then we will proceed to discuss how this techniques can be applied to Neural Networks computations. The following derivation is inspired by exercise 1 of Section 2.2 of [19].

Throughout this section we will override some of the notations that we have used so far in order to avoid cumbersome expressions.

2.3.1. Modified Dirichlet Problem

Let Ω be an open, bounded domain in $\mathbf{R}^n$, let $u\colon \Omega \to \mathbf{R}^N$, $\varpi\colon \Omega \to \mathbf{R}$ of class $C^1(\mathbf{R}^N; (0, +\infty))$ and $\mathscr{F}(u) := \int_\Omega F(x, u, \nabla u)\, dx$, then define the following functional

$$\mathscr{S}(u) := \frac{1}{2} \int_\Omega |\nabla u|^2 \, \varpi(x) dx + \mathscr{F}(u), \tag{1}$$

which is the weighted Dirichlet integral plus the $\mathscr{F}(u)$ term. We are here interested in the necessary conditions for u to be an extremizer of the *modified Dirichlet functional* $\mathscr{S}$ subject to a class of holonomic constraints. Holonomic constraints (see [19]) are subsidiary conditions of the form

$$G(x, u(x)) = 0. \tag{2}$$

When $n = 1$ and the independent variable is time if G does not depend explicitly on time we say that it is a scleronomic constraint, otherwise we will call it a rheonomic constraint.

Scalar Constraints. We start our analysis by considering a single constraint $G(u) = 0$, with $z \mapsto G(z) \in C^2(\mathbf{R}^N; \mathbf{R})$ and $|G_z|^2 \neq 0$, where we denote with G_z the gradient of the constraint.

Using the Lagrange multiplier theory we know that instead of looking for constrained stationary points of $\mathscr{S}$ we can focus on the unconstrained stationary points of the functional

$$\mathscr{S}^*(u) = \frac{1}{2} \int_\Omega |\nabla u(x)|^2 \varpi(x) - \lambda(x) G(u(x))\, dx + \mathscr{F}(u). \tag{3}$$

The constrained Euler equations therefore becomes*:

$$-\varpi \Delta u - u_{x^\alpha} \varpi_{x^\alpha} - \lambda G_z(u) + L_F(u) = 0, \tag{4}$$

where, $L_F(u)$ is the Euler operator (see [19] p. 18) and Δ is the Laplace operator.

* In this section we will make extensive use of Einstein summation convention.

Now by differentiating the constraint two times with respect to x^α, and then summing over α we obtain:

$$G_{z^i z^k} u^i_{x^\alpha} u^k_{x^\alpha} + G_{z^i} u^i_{x^\alpha x^\alpha} = 0, \tag{5}$$

or written more concisely $-G_z \cdot \Delta u = G_{z^i z^k} u^i_{x^\alpha} u^k_{x^\alpha}$. Now we can scalar multiply both sides of Eq. (4) by the vector $G_z(u)$

$$\lambda |G_z(u)|^2 = -\varpi \Delta u \cdot G_z(u) - \varpi_{x^\alpha}(u_{x^\alpha} \cdot G_z(u)) + L_F(u) \cdot G_z(u), \tag{6}$$

from where we immediately have (since we assumed $|G_z|^2 \neq 0$)

$$\lambda = \frac{\varpi G_{z^i z^k}(u) u^i_{x^\alpha} u^k_{x^\alpha} - \varpi_{x^\alpha}(u_{x^\alpha} \cdot G_z(u)) + L_F(u) \cdot G_z(u)}{|G_z(u)|^2}. \tag{7}$$

Example 1 (*Inclined Plane*). Let us consider a simple example in classical mechanics of a mass sliding on a frictionless inclined plane. The motion takes place on a two dimensional plane. This correspond to the case, $n = 1$, $N = 2$, $\varpi \equiv m$, with m a positive constant (the value of the mass). Let u^1 and u^2 be the coordinates of the mass on the plane, then the action is

$$\mathscr{S}(u) = \int_{t_0}^{t_1} \left(\frac{m}{2} |\dot{u}|^2 - mgu \cdot \hat{e}_2 \right) dt, \tag{8}$$

where $\hat{e}_2 = \binom{0}{1}$. The constraint here is $G(u) = u \cdot v$, where $v = \binom{-s}{1}$, $s > 0$. Therefore we have $G_z(u) = v$, $G_{u^i u^k} \equiv 0$ for all $i = 1, 2$, $L_F(u) = -mg\hat{e}_2$. This means that $\lambda = -mg/(1 + s^2)$ and therefore the equations of motion for the mass becomes

$$-\ddot{u}(t) + \frac{g}{1 + s^2} v - g\hat{e}_2 = 0. \tag{9}$$

In order to understand this result intuitively let us consider the rotation matrix $R(\vartheta) = \left(\begin{smallmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{smallmatrix}\right)$ and consider the change of coordinates $u \mapsto R(-\theta)u =: U$ with $\theta = \arctan s$ (the slope of the inclined plane), then the Euler equation in this new reference frame are

$$-\ddot{U}(t) - g \begin{pmatrix} \sin\theta \\ 0 \end{pmatrix} = 0, \tag{10}$$

which indeed is the well known equation of motion of a mass on an inclined plane written in the reference frame where U^1 is parallel to the plane.

Example 2 (*Pendulum*). Consider the motion of a particle on a circumference under the action of gravity (pendulum). The setting is the same as the one in Example 1; this time however the constraint is quadratic: $G(u) = \frac{1}{2}|u|^2 - \ell u \cdot \hat{e}_2$. Therefore $G_z(u) = u - \ell\hat{e}_2$ and $(G_{z^i z^j}) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, then

$$\lambda = \frac{m}{\ell^2} \big(|\dot{u}|^2 + g(\ell - u \cdot \hat{e}_2) \big). \tag{11}$$

Thus the non-linear Euler equation reads

$$-\ddot{u}(t) - \frac{|\dot{u}(t)|^2 + g(\ell - u(t) \cdot \hat{e}_2)}{\ell^2} (u(t) - \ell\hat{e}_2) - g\hat{e}_2 = 0, \tag{12}$$

this is indeed the awkward form of the non-linear equation of the pendulum $\ell\ddot{\theta} + g\sin\theta = 0$ written in the fixed reference frame with origin corresponding to the configuration $\theta = 0$.

Multiple Constraints. Consider the problem in Eq. (1) where u are subject to the constraints $G(u(x)) = 0$ for all $x \in \overline{\Omega}$ and $G(z)$ of class $C^2(\mathbf{R}^N, \mathbf{R}^r)$. Furthermore assume that the $r \times N$ matrix $G_z = (G^i_{z^j})$ has maximal rank r for all[1] $z \in \mathbf{R}^N$:

$$\operatorname{rank} G_z = r \quad \text{for all } z \in \mathbf{R}^N. \tag{13}$$

The extended functional $\mathscr{S}^*$ becomes

$$\mathscr{S}^*(u) = \frac{1}{2}\int_\Omega |\nabla u(x)|^2 \varpi(x) - \lambda_j(x) G^j(u(x))\, dx + \mathscr{F}(u) \tag{14}$$

Euler equation for this functional are $-\varpi\Delta u - u_{x^\alpha}\varpi_{x^\alpha} - \lambda_\ell G^\ell_z(u) + L_F(u) = 0$; on the other hand from the constraints one obtains $-\Delta u \cdot G^j_z(u) = G^j_{z^i z^k}(u) u^i_{x^\alpha} u^k_{x^\alpha}$ for $1 \le j \le r$. Hence $G^j_z(u)\cdot G^\ell_z(u)\lambda_\ell = \varpi G^j_{z^i z^k}(u) u^i_{x^\alpha} u^k_{x^\alpha} - \varpi_{x^\alpha}(u_{x^\alpha}\cdot G^j_z(u)) + L_F(u)\cdot G^j_z(u)$. Define $A_{j\ell}(u) := G^j_z(u)\cdot G^\ell_z(u)$, then the constrained Euler equations are

$$-\varpi\Delta u - u_{x^\alpha}\varpi_{x^\alpha} - \lambda_\ell G^\ell_z(u) + L_F(u) = 0, \tag{15}$$

with

$$\lambda_\ell = (A^{-1}(u))_{\ell j}\big(\varpi G^j_{z^i z^k}(u) u^i_{x^\alpha} u^k_{x^\alpha} - \varpi_{x^\alpha}(u_{x^\alpha}\cdot G^j_z(u)) + L_F(u) G^j_z(u)\big). \tag{16}$$

Notice that $A^{-1}(u)$ always exists because of the assumption on G_z and because $A(u)$ is a Ghram matrix and the following lemma holds:

Lemma C. *If $v_1, \ldots, v_n$ are n linear independent vectors, then the Ghram matrix $G_{ij} := (v_i, v_j)$ is positive definite.*

Proof. We have $(x, Gx) = x_i(v_i, v_j)x_j = (x_i v_i, x_j v_j) = \|v_i x_i\|^2 \ge 0$. However $\|v_i x_i\| = 0$ if and only if $x_i v_i = 0$, therefore we can conclude that $(x, Gx) > 0$ for every $x \neq 0$. ∎

2.3.2. Neural Network Constraints

The typical learning paradigm within the framework of NN consists of a model, that depends on a set of parameters W, that usually are identified as the weights of the neurons which defines the NN. In this section we are interested in a different representation of a NN in which the space of variables consist both of the weights W and the output x of the neuron and where the structure of the network is defined though a set of constrains that appropriately links together the neurons variables through the weights variables. This approach to neural networks has been originally proposed in the finite dimensional case by Yann LeCun in [34] and later developed with different declinations by [13], [55], [25] and [6]. In this section we will show how to extend the theory to functional problems. We will

[1] We could ask for a less restrictive condition here, namely that $G_z(z)$ should be full rank on all the points $z \in \mathbf{R}^N$ such that $G(z) = 0$,

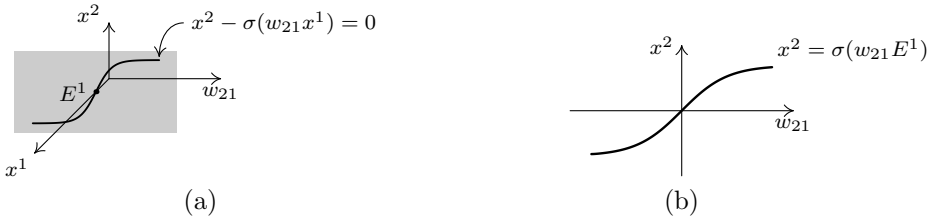


Fig. 7. Visualization of the neural constraints for the neural network $\overset{1}{\bullet}\!\!\longrightarrow\!\!\overset{2}{\bullet}$ (one input $\{1\}$ and two neurons $\{1, 2\}$). Constraint $G^2(x, W) = 0$, restricted to the plane $x^1 = E^1$, is shown in (a). In (b), such restriction is represented in the w_{21}–x^2 plane.

show that, surprisingly enough, in contrast to what happens with the standard approach of Lagrange multipliers in finite dimension, where they generally cannot be easily expressed in terms of the other variables involved in the optimization, in this case we are able to obtain an explicit representation of the multipliers.

First of all let us describe the architecture of the class of models that we will address. Given a simple digraph $D = (V, A)$ of order ν without loss of generality we can assume $V = \{1, 2, \dots, \nu\}$ and $A = \{(i, j) \in \mathbf{N}^2 \mid i \in V, j \in V\}$. A neural network constructed on D consists of a set of maps* $i \in V \mapsto x^i \in \mathbf{R}$ and $(i, j) \in A \mapsto w_{ij} \in \mathbf{R}$ together with ν constraints $G^j(x, W) = 0$ $j = 1, 2, \dots \nu$ where $(W)_{ij} = w_{ij}$. Let $\mathcal{M}_\nu(\mathbf{R})$ be the set of all $\nu \times \nu$ real matrices and $\mathcal{M}_\nu^\downarrow(\mathbf{R})$ the set of all $\nu \times \nu$ strictly lower triangular matrices over $\mathbf{R}$. If $W \in \mathcal{M}_\nu^\downarrow(\mathbf{R})$ we say that the NN has a feedforward structure. The relations $G^j = 0$ for $j = 1, \dots, \nu$ specify the computational scheme with which the information *diffuses* trough the network. In a typical network with ω inputs these constraints are defined as follows (see also Fig. 7): For any vector $\xi \in \mathbf{R}^\nu$, for any matrix $M \in \mathcal{M}_\nu(\mathbf{R})$ with entries m_{ij} and for any given C^1 map $e\colon (0, +\infty) \to \mathbf{R}^\omega$ we define the constraint on neuron j when the example $e(\tau)$ is presented to the network as

$$G^j(\tau, \xi, M) := \begin{cases} I^j(\tau, \xi) := \xi^j - e^j(\tau), & \text{if } 1 \le j \le \omega; \\ H^j(\xi, M) := \xi^j - \sigma(m_{jk}\xi^k) & \text{if } \omega < j \le \nu, \end{cases} \tag{1}$$

where $\sigma\colon \mathbf{R} \to \mathbf{R}$ is of class $C^1(\mathbf{R})$.

Our goal here is to show that such relations, that normally are considered just a local description of the compositional structure of the NN, once properly interpreted as constraints in the space $x - W$ (see Fig. 7) are suitable holonomic subsidiary conditions in the sense of Eq. 2.3.1–(13).

Like in the case of classical mechanics, when dealing with learning processes we are interested in the temporal dynamics of the variables when they are exposed to the data from which the learning is supposed to happen. For this reason in this section we can restrict ourselves to the case $n = 1$ and regard this variable as time ($x^1 = t$). Moreover because the neural constraints $G^j(x, W) = 0$

* Please notice that now x is a the variable of the variational problem, and therefore represent a mapping $t \mapsto x(t)$. It not to be intended as the independent variable of the problem described in the previous sections.

involve not only W but also x the N variables $u_1, \dots, u_N$ split into $x \in \mathbf{R}^\nu$ and $W \in \mathcal{M}_\nu(\mathbf{R})$.

Feedforward Networks. Now let us consider the case $W \in \mathcal{M}_\nu^\downarrow(\mathbf{R})$ and let us extend the theory by allowing $\mathscr{F}(x, W) := \int F(t, x, \dot{x}, \ddot{x}, W, \dot{W}, \ddot{W})\, dt$, so that, in the end, we consider the functional

$$\mathscr{A}(x, W) = \int \frac{1}{2}(|\dot{x}(t)|^2 + |\dot{W}(t)|^2)\, \varpi(t) dt + \mathscr{F}(x, W), \tag{2}$$

subject to the constraints

$$G^j(t, x(t), W(t)) = 0, \qquad 1 \le j \le \nu. \tag{3}$$

Then the following proposition holds true:

Proposition A. *The matrix* $(\frac{G_\xi}{G_M}) \in \mathcal{M}_{(\nu^2+\nu)\times\nu}(\mathbf{R})$ *is full rank.*

Proof. First of all notice that if $(G_\xi)_{ij} = G^j_{\xi^i}$ is full rank also $(\frac{G_\xi}{G_M})$ has this property. Then, since

$$G^j_{\xi^i}(\tau, \xi, M) = \begin{cases} \delta_{ij}, & \text{if } 1 \le j \le \omega; \\ \delta_{ij} - \sigma'(m_{jk}\xi^k)m_{ji} & \text{if } \omega < j \le \nu, \end{cases}$$

we immediately notice that $G^i_{\xi^i} = 1$ and that for all $i > j$ we have $G^i_{\xi^i} = 0$. This means that

$$(G^j_{\xi^i}(\tau, \xi, M)) = \begin{pmatrix} 1 & * & \cdots & * \\ 0 & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is clearly full rank. ∎

We will now discuss the updates rules (Euler-Lagrange equations) for the variables x and W derived from the stationarity conditions of the functional (2). The constrained functional is

$$\begin{aligned} \mathscr{A}^*(x, W) = \int \frac{1}{2}(m_x|\dot{x}(t)|^2 + m_W|\dot{W}(t)|^2)\varpi(t) - \lambda_j(t)G^j(t, x(t), W(t))\, dt \\ + \mathscr{F}(x, W), \end{aligned} \tag{4}$$

and its Euler equations (ELE) thus reads

$$\begin{aligned} &- m_x\varpi(t)\ddot{x}(t) - m_x\dot{\varpi}(t)\dot{x}(t) - \lambda_j(t)G^j_\xi(t, x(t), W(t)) + L^x_F(x(t), W(t)) = 0; \\ &- m_W\varpi(t)\ddot{W}(t) - m_W\dot{\varpi}(t)\dot{W}(t) - \lambda_j(t)G^j_M(t, x(t), W(t)) + L^W_F(x(t), W(t)) = 0, \end{aligned} \tag{5}$$

where $L^x_F = F_x - d(F_{\dot{x}})/dt + d^2(F_{\ddot{x}})/dt^2$, $L^W_F = F_W - d(F_{\dot{W}})/dt + d^2(F_{\ddot{W}})/dt^2$. An expression for Lagrange multiplies, as it is explained in Section is derived by differentiating two times the constraint with respect to the time and using the

obtained expression to substitute the second order terms in the Euler equations. In this case the analogue of Eq. 2.3.1–(16) is

$$\begin{aligned}\Big(\frac{G^i_{\xi^a}G^j_{\xi^a}}{m_x}+\frac{G^i_{m_{ab}}G^j_{m_{ab}}}{m_W}\Big)\lambda_j =&\varpi\big(G^i_{\tau\tau}+2(G^i_{\tau\xi^a}\dot{x}^a+G^i_{\tau m_{ab}}\dot{w}_{ab}+G^i_{\xi^a m_{bc}}\dot{x}^a\dot{w}_{bc})\\&+G^i_{\xi^a\xi^b}\dot{x}^a\dot{x}^b+G^i_{m_{ab}m_{cd}}\dot{w}_{ab}\dot{w}_{cd}\big)\\&-\dot{\varpi}(\dot{x}^aG^i_{\xi^a}+\dot{w}_{ab}G^i_{m_{ab}})+\frac{L^{x^a}_F G^i_{\xi^a}}{m_x}+\frac{L^{w_{ab}}_F G^i_{m_{ab}}}{m_W}.\end{aligned}\tag{6}$$

where G^i_τ, $G^i_{\tau\tau}$, $G^i_{\xi^a}$, $G^i_{\xi^a\xi^b}$, $G^i_{m_{ab}}$ and $G^i_{m_{ab}m_{cd}}$ are the gradients and the hessians of constraint (1) all evaluated on the trajectories $x(t)$, $W(t)$ and at the temporal instant t.

As we can expect the satisfaction of the constraints that the Euler equations (5) propagates in time must be satisfied at the beginning; and this of course imposes constraints on the way in which the system (5) can be initialized. Such conditions are discussed here below.

Initial conditions. Suppose now that we want to solve Eq. (5) with Cauchy initial conditions. Of course we must choose $W(0)$ and $x(0)$ such that $g_i(0)\equiv 0$, where we posed $g_i(t):=G^i(t,x(t),W(t))$, for $i=1,\dots,\nu$. However since the constraint must hold also for all $t\geq 0$ we must also have at least $g_i'(0)=0$. These conditions written explicitly means

$$G^i_\tau(0,x(0),W(0))+G^i_{\xi^a}(0,x(0),W(0))\dot{x}^a(0)+G^i_{m_{ab}}(0,x(0),W(0))\dot{w}_{ab}(0)=0.\tag{7}$$

Notice that the "vector" $(\dot{x}(0),\dot{W}(0))$ in order to have a motion along the constraints must be tangent to the constraints; one possible way to achieve this is to choose $\dot{x}(0)=0$ and $\dot{W}(0)=0$. This being done it remains to satisfy

$$G^i_\tau(0,x(0),W(0))=0,\tag{8}$$

which is an additional constraint on the initial conditions $x(0)$ and $W(0)$. Therefore one possible consistent way to impose Cauchy conditions for Eq. (5) is

$$\begin{cases}G^i(0,x(0),W(0))=0, & i=1,\dots,\nu;\\ G^i_\tau(0,x(0),W(0))=0, & i=1,\dots,\nu;\\ \dot{x}(0)=0;\\ \dot{W}(0)=0.\end{cases}\tag{9}$$

Higher derivative of g_i becomes automatically satisfied thanks to Eq. (5).

The following examples give a better idea of how the multipliers can be found inverting the linear equation in (6).

Example 1. We want to study only how the information that comes from the signal e is transferred to the input neurons by means of the time dependent constraint. The simplest scenario to consider is that with just one input neuron

x subject to the constraint $G = x - e$. We also just consider a free dynamic, i.e. $F = 0$. So that the only equation that we have to consider is

$$-\varpi\ddot{x} - \dot{\varpi}\dot{x} - \lambda G_x = 0, \tag{10}$$

with $\lambda = -\varpi\ddot{e} - \dot{\varpi}\dot{x}$, which means that the equation of motion for the neuron is $\ddot{x} = \ddot{e}$, which if coupled with consistent initial conditions ($x(0) = e(0)$ and $\dot{x}(0) = \dot{e}(0)$) indeed yields $x(t) = e(t)$.

Example 1. The next natural example to consider is the case of proper linear neuron, characterized by the constraint $G(t, x(t), w(t)) = x(t) - w(t)e(t)$, where e is a given input and we assume that $F = -\varpi(y-x)^2/2$, $L^x_F = \varpi(y-x)$. The ELE in this case are the following two scalar equations (we choose $\varpi(t) := \exp(\vartheta t)$)

$$\begin{aligned} \ddot{x} &= -\vartheta\dot{x} - \frac{1}{m_x}\mu + \frac{1}{m_x}(y - x); \\ \ddot{w} &= -\vartheta\dot{w} + \frac{1}{m_W}\mu e. \end{aligned} \tag{11}$$

where $\mu = \exp(-\vartheta t)\lambda$. Moreover the second derivative of the constraint gives

$$\ddot{x} - \ddot{w}e = 2\dot{w}\dot{e} + w\ddot{e}. \tag{12}$$

Multiplying the second ELE by $-e$ (which is indeed G_m) and then summing the two equations gives

$$\ddot{x} - e\ddot{w} = -\theta(\dot{x} - e\dot{w}) - \mu\frac{m_W + e^2 m_x}{m_x m_W} + \frac{1}{m_x}(y - x), \tag{13}$$

therefore

$$\mu = -\frac{m_W m_x}{m_W + e^2 m_x}\Big[2\dot{w}\dot{e} + w\ddot{e} + \vartheta(\dot{x} - e\dot{w}) - \frac{y - x}{m_x}\Big]. \tag{14}$$

Now let us define $q_0 = w$, $q_1 = x$, $q_2 = \dot{w}$, $q_3 = \dot{x}$, then

$$\dot{q}_0 = q_2, \quad \dot{q}_1 = q_3, \tag{15}$$

while

$$\begin{aligned} \dot{q}_2 &= -\vartheta q_2 - \frac{e m_x}{m_W + e^2 m_x}\Big[2q_2\dot{e} + q_0\ddot{e} + \vartheta(q_3 - eq_2)\Big] + \frac{e}{m_W + e^2 m_x}(y - q_1); \\ \dot{q}_3 &= -\vartheta q_3 + \frac{m_W}{m_W + e^2 m_x}\Big[2q_2\dot{e} + q_0\ddot{e} + \vartheta(q_3 - eq_2)\Big] + \frac{e^2}{m_W + e^2 m_x}(y - q_1). \end{aligned} \tag{16}$$

This equations can be solved with Cauchy initial conditions consistent with (9); therefore it must be true that $G_\tau(0, x(0), w(0)) = -w(0)\dot{e}(0) = 0$, so either $\dot{e}(0) = 0$ or $w(0) = 0$. This being said

$$q_0(0) = 0, \quad q_1(0) = 0, \quad q_2(0) = 0, \quad q_3(0) = 0, \tag{17}$$

is an initial condition consistent with (9).

Example 2. Consider the simplest FNN with two neurons, an input neuron and an output neuron. Then to each neuron we can assign a constraint: $G^1 = x^1(t) -$

$e(t)$, $G^2 = x^2(t) - \sigma(w_{21}(t)x^1(t))$. We also assume that $F = -\varpi(t)V(x^2(t),t) := -\varpi(t)(x^2(t) - y(t))^2/2$, where $y(t)$ is a given target. This in particular means that $L_F^x = -\varpi\binom{0}{x^2-y}$. In order to find an expression for the multipliers we need to consider Eq. (6) which is of the form $A\lambda = v$; let us now write explicitly A and v in this special case. We have that $v^1 = -(\varpi\ddot{e} + \dot{\varpi}\dot{x}^1)$, while

$$\begin{aligned} v^2 =& -2\varpi\big[\sigma''(w_{21}x^1)w_{21}x^1 + \sigma'(w_{21}x^1)\big]\dot{x}^1\dot{w}_{21} - \varpi\sigma''(w_{21}x^1)\big[w_{21}^2(\dot{x}^1)^2 \\ &+ (x^1)^2(\dot{w}_{21})^2\big] + \dot{\varpi}\big[\sigma'(w_{21}x^1)w_{21}\dot{x}^1 - \dot{x}^2 + \sigma'(w_{21}x^1)\dot{w}_{21}x^1\big] - \frac{\varpi}{m_x}(x^2 - y). \end{aligned} \tag{18}$$

The matrix A is

$$A = \begin{pmatrix} \frac{1}{m_x} & -\frac{\sigma'(w_{21}x^1)w_{21}}{m_x} \\ -\frac{\sigma'(w_{21}x^1)w_{21}}{m_x} & \frac{1+(\sigma'(w_{21}x^1))^2w_{21}^2}{m_x} + \frac{(\sigma'(w_{21}x^1))^2(x^1)^2}{m_W} \end{pmatrix}, \tag{19}$$

therefore its inverse can be written explicitly and turns out to be

$$\frac{m_x^2 m_W}{m_W + m_x\sigma'(w_{21}x^1)(x^1)^2}\begin{pmatrix} \frac{1+(\sigma'(w_{21}x^1))^2w_{21}^2}{m_x} + \frac{(\sigma'(w_{21}x^1))^2(x^1)^2}{m_W} & \frac{\sigma'(w_{21}x^1)w_{21}}{m_x} \\ \frac{\sigma'(w_{21}x^1)w_{21}}{m_x} & \frac{1}{m_x} \end{pmatrix}. \tag{20}$$

In this example we will also choose $\varpi(t) := \exp(\vartheta t)$, with $\vartheta \in \mathbf{R}$. Then if we multiply Eq. (5) by the term $\exp(-\vartheta t)$ we get

$$\begin{aligned} \ddot{x} &= -\vartheta\dot{x} - \frac{1}{m_x}\mu_j G_\xi^j - \frac{1}{m_x}\begin{pmatrix} 0 \\ x^2 - y \end{pmatrix}; \\ \ddot{W} &= -\vartheta\dot{W} - \frac{1}{m_W}\mu_j G_M^j, \end{aligned} \tag{21}$$

where $\mu_j := \exp(-\vartheta t)\lambda_j$. So that μ_j solve the same equation of λ_j multiplied by $\exp(-\vartheta t)$: $A\mu = \bar{v}$. Here $\bar{v} := \exp(-\vartheta t)v$:

$$\begin{aligned} \bar{v}^1 &= -\ddot{e} - \vartheta\dot{x}^1 \\ \bar{v}^2 &= -2\big[\sigma''(w_{21}x^1)w_{21}x^1 + \sigma'(w_{21}x^1)\big]\dot{x}^1\dot{w}_{21} - \sigma''(w_{21}x^1)\big[w_{21}^2(\dot{x}^1)^2 \\ &\quad + (x^1)^2(\dot{w}_{21})^2\big] + \vartheta\big[\sigma'(w_{21}x^1)w_{21}\dot{x}^1 - \dot{x}^2 + \sigma'(w_{21}x^1)\dot{w}_{21}x^1\big] - \frac{(x^2 - y)}{m_x}. \end{aligned} \tag{22}$$

In our case therefore Eq. (21) becomes

$$\begin{aligned} \ddot{x}^1 &= -\vartheta\dot{x}^1 - \frac{1}{m_x}\big[\mu_1 - \mu_2\sigma'(w_{21}x^1)w_{21}\big]; \\ \ddot{x}^2 &= -\vartheta\dot{x}^2 - \frac{1}{m_x}\big[\mu_2 + x^2 - y\big]; \\ \ddot{w}_{21} &= -\vartheta\dot{w}_{21} + \frac{\mu_2}{m_W}\sigma'(w_{21}x^1)x^1, \end{aligned} \tag{23}$$

where $\mu_1 = (A^{-1})_{11}\bar{v}^1 + (A^{-1})_{12}\bar{v}^2$ and $\mu_2 = (A^{-1})_{21}\bar{v}^1 + (A^{-1})_{22}\bar{v}^2$.

Now let us define $q_1^x := x^1$, $q_2^x := x^2$, $p_1^x := \dot{x}^1$, $p_2^x := \dot{x}^2$, $q^W := w_{21}$ and $p^W = \dot{w}_{21}$. Clearly, by definition, we have

$$\dot{q}^W = p^W, \quad \dot{q}_1^x = p_1^x, \quad \dot{q}_2^x = p_2^x. \tag{24}$$

Moreover, in this new notation Eq. (23)

$$\begin{aligned}
\dot{p}_1^x &= -\vartheta p_1^x - \frac{1}{m_x}\big[\mu_1 - \mu_2\sigma'(q^W q_1^x)q^W\big];\\
\dot{p}_2^x &= -\vartheta p_2^x - \frac{1}{m_x}\big[\mu_2 + q_2^x - y\big];\\
\dot{p}^W &= -\vartheta p^W + \frac{\mu_2}{m_W}\sigma'(q^W q_1^x)q_1^x,
\end{aligned} \tag{25}$$

and

$$\begin{aligned}
\bar{v}^1 &= -\ddot{e} - \vartheta p_1^x\\
\bar{v}^2 &= -2\big[\sigma''(q^W q_1^x)q^W q_1^x + \sigma'(q^W q_1^x)\big]p_1^x p^W - \sigma''(q^W q_1^x)\big[(q^W)^2(p_1^x)^2\\
&\quad + (q_1^x)^2(p^W)^2\big] + \vartheta\big[\sigma'(q^W q_1^x)q^W p_1^x - p_2^x + \sigma'(q^W q_1^x)p^W q_1^x\big] - \frac{(q_2^x - y)}{m_x}.
\end{aligned} \tag{26}$$

Eventually the inverse matrix can be written as follows

$$\frac{m_x^2 m_W}{m_W + m_x\sigma'(q^W q_1^x)(q_1^x)^2}\begin{pmatrix} \frac{1+(\sigma'(q^W q_1^x))^2(q^W)^2}{m_x} + \frac{(\sigma'(q^W q_1^x))^2(q_1^x)^2}{m_W} & \frac{\sigma'(q^W q_1^x)q^W}{m_x} \\ \frac{\sigma'(q^W q_1^x)q^W}{m_x} & \frac{1}{m_x} \end{pmatrix}. \tag{27}$$

This together with appropriate initial conditions determines the dynamic of the system.

We now show that for a particular choice of the Lagrangian the classical algorithm of Backpropagation can be recovered.

Supervised Learning and reduction to BP. In order to see how this theory can be readily applied to learning let us restrict ourselves to the case $W \in \mathcal{M}_\nu^\downarrow(\mathbf{R})$ and choose $\varpi(t) = \exp(\vartheta t)$, $\vartheta > 0$, $m > 0$. Now let us choose

$$F(t, x(t), \dot{x}(t), \ddot{x}(t), W(t), \dot{W}(t), \ddot{W}(t)) = -e^{\vartheta t}V(x(t), y(t)), \tag{28}$$

where $y(t)$ is an assigned supervision signal and

$$V(x(t), y(t)) := \frac{1}{2}\sum_{i=1}^{\eta}(y^i(t) - x^{\nu-\eta+i}(t))^2, \tag{29}$$

$x^{\nu-\eta}, \ldots, x^\nu$ being the variables associated with the outputs neurons. This choices are suggested by the analysis that we carried out in Section 2.1 and 2.2: We saw that even though the correct causal formulation of learning by means of variational method is the one that follows the De Giorgi approach, in some cases we can still use the classical action of mechanics as an "effective" functional to derive the correct differential equations.

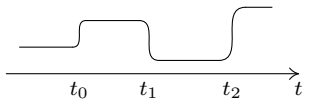
A typical input signal and the corresponding supervision signal can be constructed from a standard training set $\mathscr{L} := \{(e_\kappa, d_\kappa) \mid e_\kappa \in \mathbf{R}^\omega, d_\kappa \in \mathbf{R}^\eta, \kappa = 1, \dots, \ell\}$ in the following manner. Choose a sequence of times $\langle t_n \rangle := t_0, t_1, t_2, \dots$ such that $|t_{i+1} - t_i| =: \tau$ is constant $i \in \mathbf{N}$. Furthermore define the following sequences: $\langle E_n \rangle := e_1, \dots, e_\ell, e_1, \dots e_\ell, \dots$ and $\langle y_n \rangle := d_1, \dots, d_\ell, d_1, \dots d_\ell, \dots$. Let $R(t) := \sum_{n=0}^{\infty} \rho_\epsilon(t - t_n)$, where $\rho_\epsilon(\cdot)$ are standard Friedrichs mollifiers and define

$$\bar{E}(t) := \sum_{n=0}^{\infty} E_n \chi_{[t_{n-1}, t_n]}(t), \quad \bar{y}(t) := \sum_{n=0}^{\infty} y_n \chi_{[t_{n-1}, t_n]}(t), \tag{30}$$

where χ_A is the characteristic function of the set A and $t_{-1} = 0$. Then the signal

$$E(t) := (\bar{E} * R)(t), \quad \text{and} \quad y(t) := (\bar{y} * R)(t), \tag{31}$$

is piecewise constant signals with smooth transitions. The temporal behaviour of these signals is depicted in the side figure.



To understand the behaviour of the Euler equations (5) we observe that in the case of feedforward networks, as it is well known, the constraints $G^j(t, x, W) = 0$ can be solved for x so that eventually we can express the value of the output neurons in terms of the value of the input neurons. If we let $f^i_W(e(t))$ be the value of $x^{\nu - i}$ when $x^1 = e^1(t), \dots, x^\omega = e^\omega(t)$, then the theory defined by under subsidiary conditions is equivalent, when $m_x = 0$, to the unconstrained theory defined by

$$\int e^{\vartheta t} \Big(\frac{m_W}{2} |\dot{W}|^2 - \overline{V}(t, W(t)) \Big) \, dt \tag{32}$$

where $\overline{V}(t, W(t)) := \frac{1}{2} \sum_{i=1}^{\eta} (y^i(t) - f^i_W(E(t)))^2$. The Euler equations associated with (32) are

$$\ddot{W}(t) + \vartheta \dot{W}(t) = -\frac{1}{m_W} \overline{V}_W(t, W(t)), \tag{33}$$

that in the limit $\vartheta \to \infty$ and $\vartheta m \to \gamma$ reduces to the gradient method

$$\dot{W}(t) = -\frac{1}{\gamma} \overline{V}_W(t, W(t)), \tag{34}$$

with learning rate $1/\gamma$.

Typically the term $\overline{V}_W(t, W(t))$ in Eq. (34) can be evaluated using the Backpropagation algorithm; we will now show that Eq. (5) in the same limit used above $m_x \to 0$, $m_W \to 0$, $m_x/m_W \to 0$ reproduces Eq. (34) where the term $\overline{V}_W(t, W(t))$ explicitly assumes the form prescribed by BP. In order to see this choose $\vartheta = \gamma/m_W$ and multiply both sides of Eq. (5) and Eq. (6) by $\exp(-\vartheta t)$, then take the limit $m_x \to 0$, $m_W \to 0$, $m_x/m_W \to 0$. In this limit Eq. (5) and Eq. (6) becomes respectively

$$\begin{aligned} \dot{W} &= -\frac{1}{\gamma} \sigma'(w_{ik} x^k) \delta_i x^j; \\ G^i_{\xi^a} G^j_{\xi^a} \delta_j &= -V_{x^a} G^i_{\xi^a}, \end{aligned} \tag{35}$$

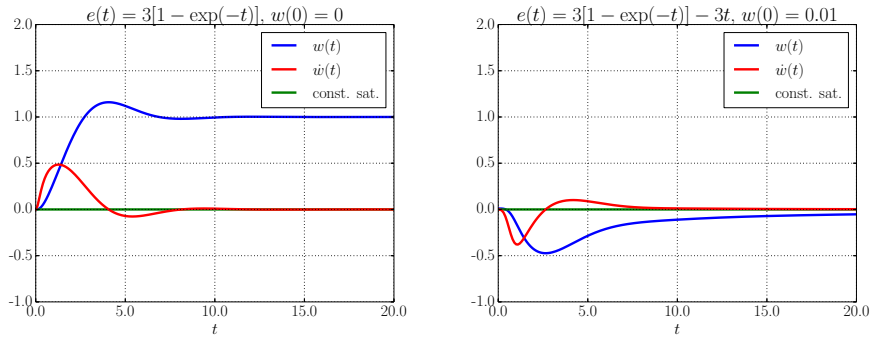


Fig. 8. Temporal evolution for the weights of a linear neuron characterized by the constraint $G = x(t) - w(t)e(t)$. We used $\overline{V} = (x-3)^2/2$ and we fixed the parameters as follows $m_x = m_W = \vartheta = 1$. The shown trajectories correspond to the initial conditions $\dot{w}(0) = \dot{x}(0) = x(0) = 0$, $w(0) = 0$ in (a), and $w(0) = 0.01$ in (b).

where δ_j is the limit of $\exp(-\vartheta t)\lambda_j$ and V_{x^a} is the a-th component of the gradient of V with respect to x. Because the matrix $G^i_{\xi^a} G^j_{\xi^a}$ not only is invertible, but it is a Gram matrix if we define $T_{ij} := G^j_{\xi^i}$, then we have $G^i_{\xi^a} G^j_{\xi^a} = (T'T)_{ij}$, with T upper triangular. Then from the second equation of (35) we have, since both T and T' are invertible:

$$T_{ij}\delta_j = -V_{x^i}. \tag{36}$$

Because T is upper triangular it is clear that this equation can be efficiently solve starting from the last line of the matrix T up to the first one. Indeed it is not hard to prove that Eq. (36) is equivalent to the backward step of the Backpropagation algorithm: Using the explicit expression for $G^j_{\xi^i}$ where i is not an input nor an output neuron we have $(\delta_{ij} - \sigma'(w_{jk}x^k)w_{ji})\delta_j = 0$ which is

$$\delta_i = \sigma'(w_{jk}x^k)w_{ji}\delta_j. \tag{37}$$

On the other hand when i is an output neuron we simply get

$$\delta_i = -V_{x^i}. \tag{38}$$

Equations (37) and (38) are exactly those that define the backward step of BP.

Simulation of the dynamics. In order to prove the soundness of the proposed theory we performed some simulations of the Euler equations and in the special case $\omega = 1$, $\eta = 1$, $\varpi = \exp(\vartheta t)$ and $F = -\exp(\vartheta t)\overline{V}(t, x(t))$, where in particular $\overline{V}(t, x(t))$ is taken to be a quadratic loss on the output neuron. To understand the learning dynamic of the weights we choose a constant supervision signal and various time-dependent input signals $e(t)$. Figure 8 shows the evolution of the weight of a single linear neuron $x(t) = w(t)e(t)$ with a target $y = 3$ and a variable input $e(t)$. In Fig. 8–(a) $e(t) \to 3$ as $t \to \infty$, and indeed $w(t)$ converges to 1. In Fig. 8–(b) $e(t) \approx 3(1-t)$ and consistently $w(t) \approx 1/(1-t)$. Notice that in

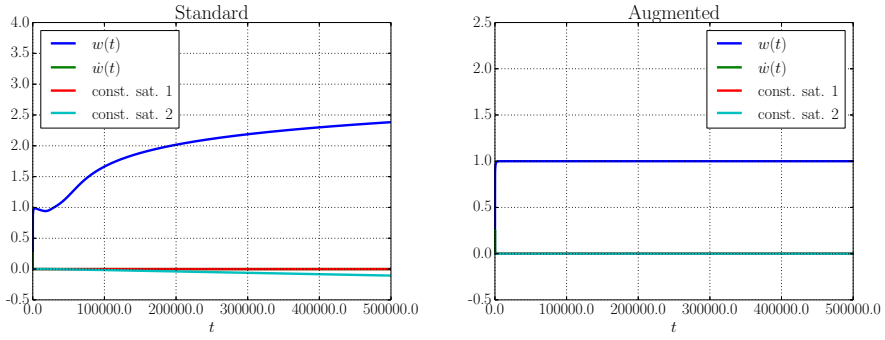


Fig. 9. Temporal evolution of the weight that connects two neurons: $G^1 = x^1 - e$, $G^2 = x^2 - \sigma(wx^1)$ with $\sigma(x) := \mathrm{Th}(x)$. We took $m_x = m_W = \vartheta = 1$ and identically null initial conditions. In (a) we considered $\overline{V} = (x^2 - \mathrm{Th}(3))^2/2$, while in (b) we chose $\overline{V} = (x^2 - \mathrm{Th}(3))^2/2 + (x^2 - \sigma(wx^1))^2/2$.

both cases the neuron constraint is always exactly satisfied. Remember that the initial conditions must be consistent with Eq. (9); in this example in Fig. 8–(a) we have $w(0) = 0$ that guaranteed $G_\tau = 0$, while in the experiment relative to Fig. 8–(b) one can choose $\dot{w}(0) \neq 0$ as the condition $G_\tau = 0$ is ensured by $\dot{e}(0) = 0$.

In Fig. 9 instead we tested the robustness of the method with respect to numerical errors by running the simulation for a longer period of time. The model here consists of two neurons NN with nonlinear activation function. We observed that due to numerical errors the system can fail to converge to the correct solution $w = 1$ (Fig. 9–(a)). This can be understood as soon as we realize that, following the ideas of Section 2.3.1, EL-equations implements only the satisfaction of the second derivative of the constraints, therefore errors on the trajectories can shift the dynamic of the system on another constraint that differs from the correct one by a linear function of time. Hence, we found that such behaviour can be effectively corrected (see Fig. 9–(b)) by adding to the potential a quadratic loss on the constraint itself.

CHAPTER THREE

VISUAL FEATURES FROM VIDEOS

Vision is the art of seeing
things invisible.
— JONATHAN SWIFT, in *Thoughts on Various Subjects; from Miscellanies* (1726)

Thou hast seen nothing yet.
— Miguel de Cervantes, in *Don Quixote, pt. I* (1605)

O! woe is me,
To have seen what I have seen, see what I see!
— WILLIAM SHAKESPEARE, in *Hamlet* (1599–1602)

Most successful computer vision algorithms work at image level, completely discarding the precious information carried by motion. In this chapter, we will discuss how the processing of visual streams naturally leads to formulate what we have called the motion invariance principle, which enables the construction of a new theory of vision that originates from variational principles. Such principled approach is well suited for a discussion on a number of interesting questions that arise in vision, and it offers a well-posed computational scheme for the discovery of convolutional filters over the retina. Differently from traditional convolutional networks, which need massive supervision, the proposed theory offers a truly new scenario for the unsupervised processing of video signals, where features are extracted in a multi-layer architecture with motion invariance. While the theory enables the implementation of novel computer vision systems, it also sheds light on the role of information-based principles to drive possible biological solutions.

For many years, the pioneering work on vision by David Marr [39], has evolved without a systematic exploration of foundations in machine learning. When the target is moved to unrestricted visual environments and the emphasis is shifted from huge labelled databases to a human-like protocol of interaction, we need to go beyond the current peaceful interlude that we are experimenting in vision and machine learning. A fundamental question a good theory is expected to answer is why children can learn to recognize objects and actions from a few supervised examples, whereas nowadays supervised learning approaches strive to achieve this task. In particular, why are they so thirsty for supervised examples? This fundamental difference seems to be deeply rooted in the different communication protocol at the basis of the acquisition of visual skills in children and machines.

So far, the semantic labeling of pixels of a given video stream has been mostly carried out at frame level. This seems to be the natural outcome of well-established pattern recognition methods working on images, which have given rise to nowadays emphasis on collecting big labelled image databases (e.g. [15]) with the purpose of devising and testing challenging machine learning algorithms. While this framework is the one in which most of nowadays state-of-the art object recognition approaches have been developing, we argue that there are strong arguments to start exploring the more natural visual interaction that animals experiment in their own environment.

This leads to process a video signal instead of image collections, that naturally leads to a paradigm-shift in the associated processes of learning to see. The idea of shifting to video is very much related to the growing interest of *learning in the wild* that has been explored in the last few years*. The learning processes that take place in this kind of environments has a different nature with respect to those that are typically considered in machine learning. Learning convolutional nets on ImageNet typically consists of updating the weights from the processing of temporally unrelated images, whereas a video carries out information when we pass through contiguous frames by smooth changes. While ImageNet is a collection of unrelated images, a video supports information only when motion is involved. In presence of fixed images that last for awhile, the corresponding stream of equal frames basically supports only the information of a single image. As a consequence, visual environments diffuse information only when motion is involved. There is no transition from one image to the next one — like in ImageNet — but, as time goes by, the information is only carried out by motion. Once we deeply capture this fundamental feature of visual environments, we early realize that we need a different theory of machine learning that must deal with video instead of a collection of independent images anymore.

A crucial problem that has been recognized by Poggio and Anselmi [45] is the need to incorporate visual invariances into deep nets that go beyond simple translation invariance that is currently characterizing convolutional networks. They propose an elegant mathematical framework on visual invariance and enlighten some intriguing neurobiological connections. Overall, the ambition of extracting distinctive features from vision poses a challenging task. While we are typically concerned with feature extraction that is independent of classic geometric transformation, it looks like we are still missing the fantastic human skill of capturing, for example, distinctive features to recognize ironed and rumpled shirts. There is no apparent difficulty to recognize shirts by keeping the recognition coherence in case we roll up the sleeves, or we simply curl them up into a ball for the laundry basket. Of course, there are neither rigid transformations, like translations and rotation, nor scale maps, that transforms an ironed shirt into the same shirt thrown into the laundry basket. Is there any natural invariance?

In this chapter, we claim that motion invariance is in fact the only invariance that we need. Translation, rotation, and scale invariance, that have been the

* See e.g. `https://sites.google.com/site/wildml2017icml/`.

subject of many studies [38], are in fact examples of invariances that can be fully gained whenever we develop the ability to detect features that are invariant under motion. Consider the simple example of your inch that moves closer and closer to your eyes. Any of its representing features that is motion invariant will also be scale invariant.

Clearly, translation, rotation, and complex deformation invariances derive from motion invariance. Humans life always experiments motion, so as the gained visual invariances naturally arise from motion invariance. Animals with foveal eyes also move quickly to focus attention on informative areas of the retina, which means that they continually experiment motion. Hence, also in case of fixed images, conjugate, vergence, saccadic, smooth pursuit, and vestibulo-ocular movements lead to acquire visual information from relative motion. We claim that the production of such a continuous visual stream naturally drives the extraction of feature that are supposed to be useful for object and action recognition. The enforcement of this consistency condition creates a mine of visual data during animal life. Interestingly, the same can happen for machines. Of course, we need to compute the optical flow at pixel level so as to enforce the consistency of all the extracted features. Early studies on this problem [28], along with recent related improvements (see e.g. [4]) suggests to determine the velocity field by enforcing brightness invariance. As the optical flow is gained, it is used to enforce motion consistency on the visual features. Interestingly, the theory we propose is quite related to the variational approach that is used to determine the optical flow in [28]. In addition to the importance of motion invariance, it is worth mentioning that an effective visual system should also develop features that do not follow such invariance. These kind of features can be conveniently combined with those that are discussed in this chapter with the purpose of carrying out high level visual tasks.

Some of the ideas presented in this chapter are somewhat inspired by the research activity reported in [22], where the authors propose the extraction of visual features as a constraint satisfaction problem, mostly based on information-based principles and early ideas on motion invariance. However, we incorporate motion invariance in the variational framework that we discussed in Chapter 2 (see also [5]), which gives rise to a time-variant differential equation, where the parameters of the model $w(t)$ which are thought as Lagrangian coordinates corresponds with the values of the convolutional filters and that we will eventually denote as $q(t)$. Unsupervised development of features from temporally coherent data has already been investigated in Slow Feature Analysis (SFA) [60, 61], with more recent applications to high-level tasks, such as action recognition [52]. The basic idea is to extract features that are "slowly varying" with respect to the "quickly varying" input signal. SFA has been applied in several contexts, and also in the case of motion estimation in video signals. Other unsupervised learning algorithms have been mostly applied to image datasets [47, 32]. More recent approaches embraces the idea of exploiting some notions of motion coherence with unsupervised learning of image-level features or with object segmentation [58, 24, 36, 43]. However to the best of our knowledge, none of the cited works

proposed a learning theory for pixel-level visual features directly formulated in the time domain and based on motion.

Motivating questions. The proposed theory offers a computational perspective on the emergence of visual features regardless of the "body" which sustains the processing. The theory is rooted on the need to address some fundamental questions that involve vision in animals, and that are likely to be very important in order to construct an effective and efficient computational model for computers. As it will become early clear, the need of visual features that support the property of motion invariance plays a central role in most of the questions outlined below.

Q1. *How can humans conquer visual skills without requiring "intensive supervision"?*

Recent remarkable achievements in computer vision are mostly based on tons of supervised examples — of the order of millions! This does not explain how can humans conquer visual skills with scarse "supervision" from the environment. Hence, there is plenty of evidence and motivations for invoking a theory strongly rooted in unsupervised learning that can be capable of explaining the emergence of features from visual data collections. While the need for theories of unsupervised learning in computer vision has been advocated in a number of papers (see e.g. [53, 35, 47, 24], so far, because of many recent successful applications, the powerful representations that arise from supervised learning, seem to attract much more interest. While information-based principles could themselves suffice to construct visual features, the absence of any feedback from the environment make those methods quite limited with respect to supervised learning. One of the founding ideas of our theory is that motion invariance inherently offers a huge amount of "free supervisions" from the visual environment, thus explaining the reason why humans do not need the massive supervision process that is dominating feature extraction in convolutional neural networks.

Q2. *How can animals gradually conquer visual skills in a truly temporal-based visual environment?*

Animals, including primates, conquer visual skills by living in their own visual environment. This is gradually achieved without needing to separate learning from test environments. At any stage of their evolution, it looks like they acquire the skills that are required to face the current tasks. On the opposite, most approaches to computer vision do not really grasp the notion of time. The typical ideas behind on-line learning do not necessarily capture the natural temporal structure of the visual tasks. Time plays a crucial role in any cognitive process. One might believe that this is restricted to human life, but more careful analyses lead us to conclude that the temporal dimension plays a crucial role in the well-positioning of most challenging cognitive tasks, regardless of whether they are supported by humans or machines. Nowadays dominating trend leads to struggle for the acquisition

of huge labeled databases, while the truly incorporation of time might led to a paradigm shift in the interpretation of the learning and test environment and construct visual features without needing any labeling. The theory proposed here is framed in the context of agent life characterized by the ordinary notion of time, which emerges in all its facets. We are not concerned with huge supervised visual data repositories, but merely with the agent life in its own visual environments. The extraction of features in such a temporal-based visual environment is the main objective of this chapter.

Q3. *Can animals see in a world of shuffled frames?*

One might figure out what human life could have been in a world of visual information with shuffled frames. Could children really acquire visual skills in such an artificial world, which is the one we are presenting to machines? Notice that in a world of shuffled frames, for a video to be recorded, we require a space that is significantly larger than the space required to store the corresponding temporally coherent visual stream. This is a serious warning that is typically neglected. As a consequence, any recognition process is likely to be remarkably more difficult when shuffling frames, which clearly indicates the importance of keeping the spatio-temporal structure that is offered by nature. This calls for the formulation of a theory of learning capable of capturing spatiotemporal structures. Basically, we need to abandon the indisputable issue of restricting computer vision to the processing of images. The reason for formulating a theory of learning on video instead of on images is not only rooted in the curiosity of grasping the computational mechanisms that take place in nature. It looks like that, while ignoring the crucial role of temporal coherence, learning visual features leads to tackling a problem that is remarkably more difficult than the one nature has prepared for humans! We conjecture that animals could not see in a world of shuffled frames, which indicates that such an artificial formulation might led to a very hard problem. In a sense, the very good results that we already can experiment nowadays on the extraction of visual features are quite surprising, but they are mostly due to the stress of the computational power and the artificial framework of supervised learning. The theory proposed in here relies on the choice of capturing temporal structures in natural visual environments, which is claimed to simplify dramatically the problem at hand, and to give rise to a reduce dramatically the computational burden.

Q4. *How can humans attach semantic labels at pixel level?*

Humans provide scene interpretation thanks to linguistic descriptions. This requires a deep integration of visual and linguistic skills, that are required to come up with compact, yet effective visual descriptions. However, amongst these high level visual skills, it is worth mentioning that humans can attach semantic labels to a single pixel in the retina. While this decision process is inherently interwound with a certain degree of ambiguity, it is remarkably effective. The linguistic attributes that are extracted are related to the

context of the pixel that is taken into account for label attachment, while the ambiguity seems to be mostly a linguistic more than a visual issue. The proposed theory addresses directly this visual skill since the hidden labels can be extracted for a given pixel at different levels of abstraction.
The bottom line is that human-like linguistic descriptions of visual scenes is gained on top of pixel-based feature descriptions that, as a byproduct, must allow us to perform semantic labeling. Interestingly, there is more; as it will be shown in the following, there are in fact computational issues that lead us to promote the idea of carrying out the feature extraction process while focussing attention on salient pixels.

Q5. *What could drive the functional difference between the ventral and dorsal mainstream in the visual cortex?*

It has been pointed out that the visual cortex of humans and other primates is composed of two main information pathways that are referred to as the ventral stream and dorsal stream [20]. The ventral "what" and the dorsal "where/how" visual pathways are traditionally distinguished, so as the ventral stream is devoted to perceptual analysis of the visual input, such as object recognition, whereas the dorsal stream is concerned with motion ability in the interaction with the environment. The enforcement of motion invariance is clearly conceived for extracting features that are useful for object recognition to assolve the "what" task. Of course, neurons with built-in motion invariance are not adequate to make spatial estimations. The model behind the learning of the filters indicates the need to access to velocity estimation, which is consistent with neuroanatomical evidence. Interestingly, we will see that the theory also advocates the need of hierarchical structures for the dorsal mainstream, but there is one more reason for those structures in the ventral stream.

Q6. *Why do we need a hierarchical architecture with receptive fields?*

Beginning from early studies by Hubel and Wiesel [30], neuroscientists have gradually gained evidence that the visual cortex presents a hierarchical structure, and that the neurons process the visual information on the basis of inputs restricted to receptive field. Is there any reason why this solution has been developed? We can promptly realize that, even though the neurons are restricted to compute over receptive fields, deep structures easily conquer the possibility of taking large contexts into account for their decision. Is this biological solution driven by computational laws of vision? We provide evidence of the fact that receptive fields do favor the acquisition of motion invariance which, as already stated, is the fundamental invariance of vision. Since hierarchical architectures is the natural solution for developing more abstract representations by using receptive fields, it turns out that motion invariance is in fact at the basis of the biological structure of the visual cortex. The computation at different layers yields features with progressive degree of abstraction, so as higher computational processes are expected to use all the information extracted in the layers.

Q7. *Why do animals focus attention?*

The retina of animals with well-developed visual system is organized in such a way that there are very high resolution receptors in a restricted area, whereas lower resolution receptors are present in the rest of the retina. this convenient? One can easily argue that any action typically takes place in a relatively small zone in front of the animals, which suggests that the evolution has led to develop high resolution in a limited portion of the retina. On the other hand, this leads to the detriment of the peripheral vision, that is also very important. In addition, this could apply for the dorsal system whose neurons are expected to provide information that is useful to support movement and actions in the visual environment. At a first glance, the ventral mainstream, with neurons involved in the "what" function, does not seem to benefit from foveal eyes. The theory proposed here strongly supports the need for foveal retinas, when we need to achieve an efficient construction of visual features delegated to sustain object recognition. However, it will be argued that the most important reason for focussing attention is that of dramatically simplifying the computation and limit the ambiguities that come from the need to sustaining a parallel computation over each frame.

Q8. *Why do foveal animals perform eye movements?*

Human eyes make jerky saccadic movements during ordinary visual acquisition. One reason for these movements is that the fovea provides high-resolution in portions of about $1,2$ degrees. Because of such a small high resolution portions, the overall sensing of a scene does require intensive movements of the fovea. Hence, the foveal movements do represent a good alternative to eyes with uniformly high resolution retina. On the other hand, the preference of the solution of foveal eyes with saccadic movements is arguable; while a uniformly high resolution retina is more complex to achieve than foveal retina, saccadic movements in this case are less important. The information-based theory presented here makes it possible to conclude that foveal retina with saccadic movements is in fact a solution that is computationally sustainable and very effective.

Q9. *Why does it take 8-12 months for newborns to achieve adult visual acuity?*

There are surprising results that come from developmental psychology on what a newborn see. Charles Darwin came up with the following remark:

> *It was surprising how slowly he acquired the power of following with his eyes an object if swinging at all rapidly; for he could not do this well when seven and a half months old.*

At the end of the seventies, this early remark was given a technically sound basis [17]. In the paper, three techniques — optokinetic nystagmus (OKN), preferential looking (PL), and the visually evoked potential (VEP) — were used to assess visual acuity in infants between birth and 6 months of age. More recently, the survey by Braddick and Atkinson [11] provides an in-depth discussion on the state of the art in the field. It is clearly stated that

for newborns to gain adult visual acuity, depending on the specific visual test, several months are required. Is the development of adult visual acuity a biological issue or does it come from higher level computational laws? We have already commented in Section 2.2.3 that the "speed" at which data comes interacts with the natural dynamics of learning when the process is modelled through a differential equation. Therefore we can argue that the blurring process taking place in newborns is in fact a natural strategy to solve online learning problems under causality requirements. Moreover, the strict limitations both in terms of spatial and temporal resolution of the video signal, according to the theory, help conquering visual skills.

Q10. *Causality and Non Rapid Eye Movements (NREM) sleep phases*

Computer vision is mostly based on huge training sets of images, whereas humans use video streams for learning visual skills. Notice that because of the alternation of the biological rhythm of sleep, humans somewhat process collections of visual streams pasted with relaxing segments composed of "null" video signal. This happens mostly during NREM phases of sleep, in which also eye movements and connection with visual memory are nearly absent. Interestingly, the Rapid Eye Movements (REM) phase is, on the opposite, similar to ordinary visual processing, the only difference being that the construction of visual features during the dream is based on the visual internal memory representations [2]. As a matter of fact, the process of learning the filters experiments an alternation of visual information with the reset of the signal. We provide evidence to claim that such a relaxation coming from the reset of the signal nicely fits the overall objective of the visual agent.

In particular, throughout this chapter, we will see that the reset of the visual information favors the optimization under causality requirements. From one side it provides a good way to enforce relaxation of the derivatives of the filters, while on the other it provides a natural way to perform input segmentation as described in Section 2.2.2. Hence, the theory offers an intriguing interpretation of the role of eye movement and of sleep for the optimal development of visual features. In a sense, the theory also offers a general framework for interpreting the importance of the day-night rhythm in the development of visual features.

This chapter is organized as follows: Section 3.1 and 3.2 contains the description of the theory both formulated on a continuous and a discrete retina. In Section 3.2 a first analysis of causality is presented based on the ideas of relaxation and reset briefly introduced in Section 2.2.2. In the last section of the chapter we will revisit the theory addressing the causality issue in a more systematic following the analyses presented in Section 2.2.

We decided to present both approaches in order to reinforce the analysis of Chapter 2 that showed how causality issues are central in this kind of discussions.

3.1. GENERAL FORMULATION OF THE PROBLEM

We are given a retina $\Sigma \subset \mathbf{R}^2$, which can formally be regarded as a compact subset of the plane; for the moment we will not assume any specific shape. The purpose of this chapter is that of analyzing the mechanisms that give rise to the construction of local features for any pixel $x \in \Sigma$ of the retina, at any time t. These features, along with the video itself, can be regarded as visual fields, that are defined on the retina and on a given time horizon $[0, T]$. As it will be clear in the remainder of the chapter, a set of symbols are extracted at any layer of a deep architecture, so as any pixel — along with its context — turns out to be represented by the list of symbols extracted at each layer. The computational process that we define involves the video as well as appropriate vector fields that are used to express a set of pixel-based features properly used to capture contextual information. The video, as well as all the involved fields, are defined on the parallelepiped $\Pi = \Sigma \times [0, T]$. In what follows, points on the retina will be represented with two dimensional vectors $x = (x_1, x_2)$ on a defined coordinate system on the retina. The temporal coordinate is usually denoted by t, and, therefore, the video signal on the pair (x, t) is $C(x, t)$. The color field can be thought of as a special field that is characterized by the RGB color components of any single pixel; in this case $m = 3$.

Now, we are concerned with the problem of extracting visual features that, unlike the components of the video, express the information associated with the pair (x, t) and with its spatial context. Basically, one would like to extract visual features that characterize the information in the neighborhood of pixel x. A possible way of constructing this kind of features is to construct the map*

$$\Phi_i(x,t) = \sum_{j=1}^{m} \int_{\Sigma} \varphi_{ij}(x, y, t, C_j(y,t))\, dy. \tag{1}$$

Here, the feature defined by index $i = 1, \ldots, n$, that is denoted by $\Phi_i(x, t)$ presents a spatial dependence on any pixel $y \in \Sigma$. Here we assume that n symbols are generated from the m components of the video. In the special case in which such a dependence only involves the distance from the pixel of coordinates x on which we want to determine the feature, the above equation reduces to

$$\Phi_i(x,t) = \sum_{j=1}^{m} \int_{\Sigma} \varphi_{ij}(x - y, t, C_j(y,t))\, dy, \tag{2}$$

which becomes the convolutional computation in case of linear filters φ_{ij}, that is

$$\Phi_i(x,t) = \sum_{j=1}^{m} \int_{\Sigma} \varphi_{ij}(x - y, t) C_j(y,t)\, dy \tag{3}$$

Notice that $\varphi(z, t)$ is responsible of expressing the spatial dependencies, and that one could also extend the context in the temporal dimension. However,

* In the rest of the chapter when the expressions will become more involved we will make extensive use of Einstein summation convention.

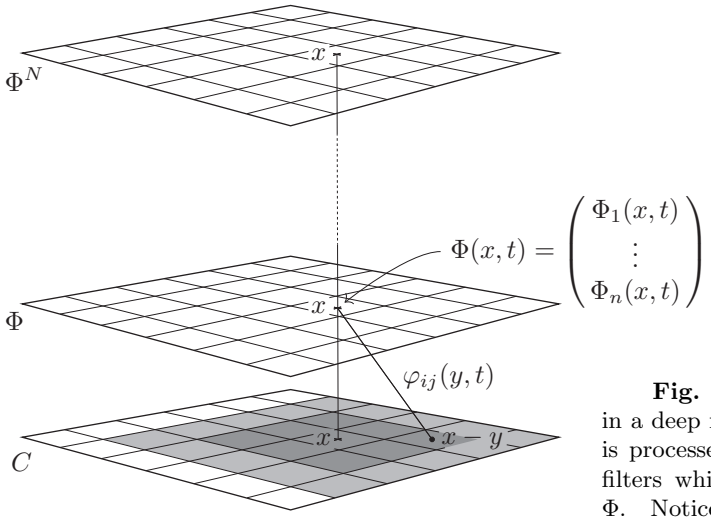


Fig. 10. Computation in a deep network. The input is processed by convolutional filters which transform C to Φ. Notice that the features are extracted at different levels on the same pixel x.

the immersion in the temporal dimension that arises from the formulation given here makes it reasonable to begin restricting the contextual information to spatial dependencies on the the retina. In addition, it is worth mentioning that the agent is expected to return a decision also in case of fixed images, which represents a further element for considering features as defined in Eq. (3).

In general, the kernel φ can be regarded as a map from $\Sigma \times \Sigma \times [0,T] \to \mathbf{R}^{n,m}$. Whenever $\varphi(x,y,t) \to \varphi(x-y,t)$ the above definition reduces to an ordinary spatial convolution. Notice that while the kernel $\varphi(x,y,t)$ can handle the ambiguities that arise from the the presence of strong visual deformations of the same features in the same frame at time t, the same does not hold for $\varphi(x-y,t)$, that only reasonably deals with those deformations while focusing attention on x at time t. This issue will be widely covered in the following, but it is already clear that the convolutional filter $\varphi(x-y,t)$ can face strong visual deformation only when supported by focus of attention driven computation. The presence of multiple deformations in the same frame yields inconsistent decisions, so as only an "averaging solution" can be discovered. The computation of $\Phi(x,t)$ yields a field with n features, instead of the three components of color in the video signal. However, Eq. (3) can be used for carrying out a piping scheme where a new set of features Φ^2 is computed from Φ and so forth (see Fig. 10). Of course, this process can be continued according to a deep computational structure with a homogeneous convolutional-based computation, which yields the features at the p convolutional layer. The theory proposed here focuses on the construction of any of these convolutional layers which are expected to provide higher and higher degree of abstraction as we increase the number of layers. The *filters* φ completely determine the features $\Phi(x,t)$. In what follows we will formulate a theory for the discovery of φ that is based on three driving principles:

1. *Optimization of information-based indices:* We use an information-based approach to determine φ. Beginning from the color field C, we attach symbol $y_i \in \Sigma$ of a discrete vocabulary to pixel (x,t) with probability $\Phi_i(x,t)$. The principle of Maximum Mutual Information (MMI) is a natural way of maximizing the transfer of information from the visual source, expressed in terms of mixtures of colors, to the source of symbols $y_i \in \Sigma$. Clearly, the same idea can be extended to any layer in the hierarchy. Once we are given a certain visual environment over a certain time horizon $[0,T]$ — which can be extended to $[0,+\infty)$ — once the filters φ have been defined, the mutual information turns out to be a functional of φ, that is denoted as $\mathcal{I}(\varphi)$. However, in the following, it will be shown that the more general view behind the the maximum entropy principle (MaxEnt) offers a better framework for the formulation of the theory.
2. *Motion invariance:* While information-based indices optimize the information transfer from the input source C to the symbols, the major cognitive issues of invariances are not covered. The same object, which is presented at different scales and under different rotations does require different representations, which transfers all the difficulty of learning to see to the subsequent problems interwound with language interpretation. Hence, it turns out that the most important requirement that the visual field Φ must fulfill is that of exhibiting the typical cognitive invariances that humans and animals experiment in their visual environment. We claim that there is only one fundamental invariance, namely that of producing the same representation for moving pixels. This incorporates classic scale and rotation invariances in a natural way, which is what is experimented in newborns. Objects comes at different scale and with different rotations simply because children experiment their movement and manipulation. As we track moving pixels, we enforce consistent labeling, which is clearly far more general than enforcing scale and rotation invariance. We claim that the enforcement of motion constraint is the key for the construction of a truly natural invariance.
3. *Parsimony principle:* Like any principled formulation of learning, we require the filters to obey the parsimony principle. Amongst the philosophical implications, it also favors the development of a unique solution. The development of filters that are consistent with the above principles requires the construction of an on-line learning scheme, where the role of time becomes of primary importance. The main reason for such a formulation is the need of imposing the development of motion invariance features. Given the filters φ, there are two parsimony terms, one $\mathcal{P}(\varphi)$, that penalizes abrupt spatial changes, and another one, $\mathcal{K}(\varphi)$ that penalizes quick temporal transitions.

Overall, the process of learning is regarded as the minimization of the *cognitive action*

$$\mathcal{A}(\varphi) = -\mathcal{I}(\varphi) + \lambda_M \mathcal{M}(\varphi) + \lambda_P \mathcal{P}(\varphi) + \lambda_K \mathcal{K}(\varphi), \tag{4}$$

where $\lambda_M, \lambda_P, \lambda_K$ are positive multipliers. While the first and third principles are typically adopted in classic unsupervised learning, motion invariance does

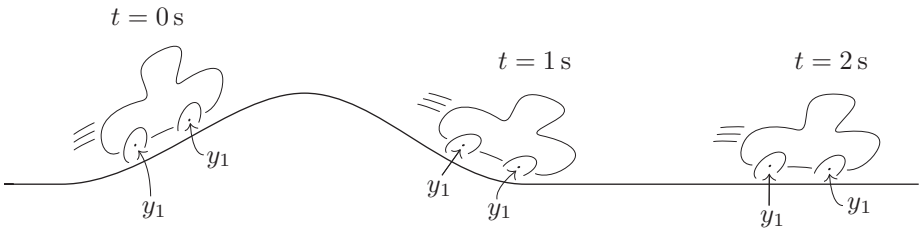


Fig. 11. Motion invariance in the feature extraction process. The symbol y_1, that defines a features at the beginning of motion ($t = 0$ s), must be coherently extracted during the movement — see the enforcement of the coherence requirement at $t = 1, 2$ s.

characterize our approach. Of course, there are visual features that do not obey the motion invariance principle. Animals easily estimate of the distance to the objects in the environment, a property that clearly indicates the need for features whose value do depend on motion. The perception of vertical visual cues, as well as a reasonable estimation of the angle with respect to the vertical line also suggests the need for features that are motion dependent. Since the above action functional $\mathcal{A}(\varphi)$ depends on the choice of the multipliers $\lambda_M, \lambda_P, \lambda_K$, it is quite clear that there is a wide range of different behavior that depend on the relative weight that is given to the terms that compose the action. As it will be shown in the following, the minimization of $\mathcal{A}(\varphi)$ can be given an efficient computational scheme only if we give up to optimize the information transfer in one single step and rely on a piping scheme that clearly reminds deep network architectures.

Now, we provide arguments to support the principled framework of this approach. Like for human interaction, visual concepts are expected to be acquired by the agents solely by processing their own visual stream along with human supervisions on selected pixels, instead of relying on huge labelled databases. In this new learning environment based on a video stream, any intelligent agent willing to attach semantic labels to a moving pixel is expected to take coherent decisions with respect to its motion. Basically, any label attached to a moving pixel has to be the same during its motion. Hence, video streams provide a huge amount of information just coming from imposing coherent labeling, which is likely to be the primary information associated with visual perception experienced by any animal. Roughly speaking, once a pixel has been labeled, the constraint of coherent labeling virtually offers tons of other supervisions, that are essentially ignored in most machine learning approaches working on big databases of labeled images. It turns out that most of the visual information to perform semantic labeling comes from the motion coherence constraint, which might explain the reason why children learn to recognize objects from a few supervised examples. The linguistic process of attaching symbols to objects takes place at a later stage of children development, when he has already developed strong pattern regularities. We conjecture that, regardless of biology, the enforcement of motion coherence constraint is a high level computational principle that plays a fundamental role for discovering pattern regularities.

Concerning the MMI principle, it is worth mentioning that it can be regarded as a special case of the MaxEnt principle when the constraints correspond with the soft-enforcement of the conditional entropy, where the weight of its associated penalty is the same as that of the entropy (see e.g. [40]. Notice that while the maximization of the mutual information nicely addresses the need of maximizing the information transfer from the source to the selected alphabet of symbols, it does not guarantee temporal consistency of this attachment. Basically, the optimization of the index is also guaranteed by using the same symbol for different visual cues. Motion consistency faces this issue for any pixel, even if it is fixed. As for the adoption of the parsimony principle in visual environments, we can use appropriate functionals to enforce both the spatial and temporal smoothness of the solution. While the spatial smoothness can be gained by penalizing solutions with high spatial derivatives — including the zero-order derivatives — temporal smoothness arises from the introduction of kinetic energy terms which penalizes high velocity and, more generally, high temporal derivatives.

The agent behavior turns out to be driven by the minimization of an appropriate functional that combines the all above principles. The main result is that this optimization can be interpreted in terms of laws of nature expressed by a temporal differential equation.

Basically, the process of learning consists of determining

$$\hat{\varphi} = \arg\min_{\varphi} \mathcal{A}(\varphi).$$

We will show how we can get the filters $\hat{\varphi}$ by addressing the problem of determining stationary points of the action $\mathcal{A}$ and, moreover, we discuss the existence of such filters $\hat{\varphi}$. The filters are determined by imposing

$$\delta\mathcal{A}(\varphi) = 0, \tag{5}$$

that is the nullification of the variation of the action, which corresponds with the stationarity condition on $\mathcal{A}$. It is worth mentioning that this does not correspond with the classic gradient flow used in machine learning, since in that case the filters are updated by using the gradient heuristics towards the stationary condition. The consequences of imposing condition (5) is mostly discussed in Section 3.1.2, where we prove that, when considering the continuous setting of computation in which $\varphi_{ij}(z,t)$ are the unknown filters, there is no local solution to this problem, since any stationary point of this functional turns out to be characterized by a integro-differential equation. Interestingly, we show that we can naturally gain a local solution when introducing the classic notion of receptive field. This issue turns out to be relevant also in case we deal with a discrete retina. The equations contain coefficients which inherits by the time-variance from the video. The analysis carried out shows how can we attack the problem either in the case in which the agent is expected to learn from a given video stream with the purpose to work on subsequent test collections, or in the case in which the agent lives in a certain visual environment, where there is no

distinction between learning and test phases. Basically, it is pointed out that only the second case leads to a truly interesting and novel result.

The puzzle of extracting robust cues from visual scenes has only been partially faced by nowadays successful approaches to computer vision. The remarkable achievements of the last few years have been mostly based on the accumulation of huge visual collections gathered by crowdsourcing. An appropriate set up of convolutional networks trained in the framework of deep learning has given rise to very effective internal representations of visual features. They have been successfully used by facing a number of relevant classification problems by transfer learning. Clearly, this approach has been stressing the power of deep learning when combining huge supervised collections with massive parallel computation. We argue that while stressing this issue we have been facing artificial problems that, from a pure computational point of view, are likely to be significantly more complex than natural visual tasks that are daily faced by animals. In humans, the emergence of cognition from visual environments is interwound with language. This often leads to attack the interplay between visual and linguistic skills by simple models that, like for supervised learning, strongly rely on linguistic attachment. However, when observing the spectacular skills of the eagle that catches the pray, one promptly realizes that for an in-depth understanding of vision, that likely yields also an impact in computer implementation, one should begin with a neat separation with language! This theory is mostly motivated by the curiosity of addressing a number of questions that arise when looking at natural visual processes. While they come from natural observation, they are mostly regarded as general issues strongly rooted in information-based principles, that we conjecture are of primary importance also in computer vision.

3.1.1. A Functional for Vision

We can provide an interpretation of the processing carried out by our visual agent in the framework of information theory. The basic idea is that the agent produces a set of symbols from a given alphabet while processing the video. Unlike traditional approaches to computer vision, we begin considering that maps on the retina are refined with the final purpose of transforming the color field, which reports pixel-based information, into visual features that take the pixel context into account. As such, one could expect each pixel to be associated with a remarkable number of features that somehow express the visual information in its neighborhood. A similar map of features, $\Phi(x,t)$ is clearly reporting an enriched color field that, just like $C(x,t)$, still operates at pixel level. In doing so, all subsequent cognitive tasks that relies on video can benefit of the processing on $\Phi(x,t)$ that, unlike $C(x,t)$, is expected to express relevant visual features that emerge from the context. It will be shown that the search for an appropriate enrichment of the color field leads to important architectural conclusions that address some of questions raised in the previous section and very much support nowadays emphasis on the deep networks.

MMI principle. The purpose of the visual agent is to generate symbols from the video. We will make use of the Maximum Mutual Information principle (MMI), according to which we want to maximize the transfer of information from the input to the generated symbols. As it will be shown later, this can also be reformulated within the framework of the Maximum Entropy principle [31].

Let us define random variables X and T, which take into account the spatiotemporal probability distribution, while Y is used to specify the probability distribution over the possible symbols, and F to specify the video frame. Basically, the realization of these of (X, T, F) is the triple (x, t, f), which describes the spatiotemporal pair (x, t) (pixel-time) at frame f, that is clearly characterized by the given video signal at time t. In order to assess the information transfer from (X, T, F) to Y we consider the corresponding mutual information I. Clearly, it is zero whenever random variable Y is independent of X, T and F. The mutual information can be expressed by

$$I(Y; X, T, F) = S(Y) - S(Y \mid X, T, F). \tag{1}$$

The conditional entropy $S(Y \mid X, T, F)$ is given by

$$S(Y \mid X, T, F) = -\int_{\Omega} \sum_{i=1}^{n} dP_{X,T,F}\ p_i \log p_i \tag{2}$$

where p_i is the probability of Y conditioned to the values of X, T and F, $dP_{X,T,F}$ is the joint measure of the variable (X, T, F), and Ω is a Borel set in the (X, T, F) space. The agent is supposed to generate symbols $y_i,\ i = 1, \ldots, n$ along with the corresponding probabilities. Now, let us make two fundamental assumptions:

1. The conditional probability p_i is given by the i-th feature field $\Phi_i(x, t)$. Notice that one can also distinguish between the feature map $\Phi_i(x, t)$ and the symbols to be used in the codebook. In that case, we need an additional map $\Phi(x, t) \to \Psi(\Phi(x, t))$, that could be properly expressed by a feedforward neural network that is charged of computing the probability p_i.
2. Random variables X, T, F follows an ergodic-like assumption, so as we can perform the replacement:

$$\int_{\Omega} dP_{X,T,F} \longrightarrow \int_{\Pi} d\mu. \tag{3}$$

A reasonable measure is given by $d\mu = f(x, t)\, dxdt$, where $dxdt$ is the Lebesgue measure on Π; basically, this comes from the visual environment on which the agent is supposed to operate. Furthermore, we will assume that we are given the trajectory of the focus of attention $t \mapsto a(t) \in \Sigma$ and that $f(x, t)$ is factorized according to

$$f(x, t) = g(x - a(t))\varpi(t), \tag{4}$$

This ergodic-like translation of the probabilistic measure has been chosen in such a way that the density is higher where the eye is focussing attention, that is in the neighborhood of $a(t)$; this can be achieved by means of a function $g(x - a(t))$ peaked on the focus of attention. As it will be shown in the following, and as we

have already commented in Section 2.1 the weight factor $\varpi(t)$ plays a crucial role in the establishment of dissipation and causality, two properties indeed related to the enforcement of a temporal direction.

Notice that in truly active environments humans and robots can select even the environment, which may result in a remarkable variability of the probability distributions. For instance, living like Eskimos leads to acquire visual information that are remarkable different from Newyorkèse. Regardless of the huge visual environmental gap, however, humans seem to adapt very well their visual system when moving from New York to snow territories and vice versa. This suggests that when learning in natural environments focus of attention strategies, that are associated with the computation $a(t)$, seem to be remarkably important in the acquisition of visual skills.

The research on focussing of attention trajectories $a(t)$ is rooted on solid studies at the crossroad of neuroscience and computer vision, and it has been recently given a formulation [62, 63] that is very much aligned with the theoretical framework of this work.

Whenever assumptions 1. and 2. holds, we can rewrite the conditional entropy defined by Eq. (2) as

$$S(Y \mid X, T, F) = -\int_{\Pi} \sum_{i=1}^{n} \Phi_i(x,t) \log \Phi_i(x,t)\, d\mu(x,t). \tag{5}$$

Similarly for the entropy of the variable Y we can write

$$S(Y) = -\sum_{i=1}^{n} \Pr(Y = y_i) \log \Pr(Y = y_i). \tag{6}$$

Now, if we use the law of total probability to express $\Pr(Y = y_i)$ in terms of the conditional probability p_i and use the above assumptions we get

$$\Pr(Y = y_i) = \int_{\Omega} dP_{X,T,F}\ p_i = \int_{\Pi} \Phi_i(x,t)\, d\mu(x,t). \tag{7}$$

Then

$$S(Y) = -\sum_{i=1}^{n} \Big(\int_{\Pi} \Phi_i(x,t)\, d\mu(x,t) \Big) \log \Big(\int_{\Pi} \Phi_i(x,t) d\mu(x,t) \Big). \tag{8}$$

Finally the mutual information becomes

$$I(Y; X, T, F) = \sum_{i=1}^{n} \Big(\int_{\Pi} \Phi_i \log \Phi_i\, d\mu - \int_{\Pi} \Phi_i\, d\mu \log \int_{\Pi} \Phi_i\, d\mu \Big). \tag{9}$$

Of course, $\forall x, t$ the field $\Phi_i(x,t)$ is subject to the probabilistic constraints

$$\begin{array}{ll} \sum_i \Phi_i(x,t) = 1 & \text{(normalization)} \\ 0 \le \Phi_i(x,t) \le 1 & \text{(positivity)} \end{array} \tag{10}$$

In the case there is an additional neural map Ψ to determine the probability, the normalization is moved to the range of the map itself, which allows the typical presence of more distributed representations on Φ.

MaxEnt principle. An agent driven by the MMI principle carries out an unsupervised learning process aimed at discovering the symbols defined by random variable Y. Interestingly, when the constraints are given a soft-enforcement, the MMI principle has a nice connection with the Max-Ent principle [31]: The maximization of the mutual information is somewhat related to the maximization of the entropy while softly-enforcing the constraint that the conditional entropy is null. In particular, in MMI both the entropy terms get the same value of the weight, but one can think of different implementations of the MaxEnt principle that very much depend on the choice of the weights of the two entropy terms. As an extreme case, one can also remove the conditional entropy term and consider motion invariance only. The satisfaction of the conditional entropy constraint needs to be paired with the maximization of the entropy, which protects us from the development of trivial solutions (see [21] pp. 99–103 for further details). Of course, the probabilistic normalization constraints stated by Eq. (10) comes along with the information-based formulation. While the computational mechanism that drives the discovery of the symbols described here is inspired by MaxEnt, a well-posed learning process requires that the map which originates the symbols be subjected to some kind of parsimony assumption. The conditional entropy constraint only involves the value taken by Φ_i which depends on $\varphi_{ij}(x,t)$, but there is no structural enforcement on the function φ_{ij}; its spatiotemporal changes are ignored. Ordinary regularization issues suggest to select functions φ_{ij} such that $\|P_{x,t}\varphi_{ij}\|_{L^2}$ is small, where $P_{x,t}$ is a spatiotemporal differential operator. A simplified, yet effective choice is that of separating the spatial from the temporal regularization and consider

$$\frac{\lambda_P}{2}\int_\Pi dtdx\,\varpi(t)(P_x\varphi_{ij}(x,t))^2+\frac{\lambda_K}{2}\int_\Pi dtdx\,\varpi(t)(P_t\varphi_{ij}(x,t))^2, \tag{11}$$

is "small", where P_x, P_t are spatial and temporal differential operators, and λ_P, λ_K are non-negative reals*. Notice that the ergodic-like translation of $d\mu$, in this case, only involves the temporal factor $\varpi(t)$.

Second, as already pointed out, the visual features that in the ventral mainstream are involved in the "what" function need to be motion invariant. Just like an ideal fluid is adiabatic — meaning that the entropy of any particle fluid remains constant as that the particles move about in space — in a video, once we have assigned the correct symbol to a pixel, it must be conserved as the pixel moves on the retina. If we focus attention on a the pixel x at time t, which moves according to the trajectory $x(t)$ then this is formally stated by $\Phi_i(x(t),t)=c$, being c a constant. This "adiabatic" condition is thus expressed by the condition $d\Phi_i/dt=0$, which yields

$$\partial_t\Phi_i+\dot{x}_j\partial_j\Phi_i=0, \tag{12}$$

where ∂_k is the partial derivative with respect to x_k. In a video we can assign to each pixel x at a certain instant t the velocity vector $v(x,t)$ (from the optical

* A simple introduction to differential operators that are appropriate in this context is given in [21], pp. 512–516.

flow). Therefore we can turn the condition (13) into a condition for each pixel x at a certain time t:

$$\partial_t \Phi_i(x,t) + v_j(x,t)\partial_j \Phi_i(x,t) = 0; \tag{13}$$

we will assume that the velocity field $(x,t) \mapsto v(x,t) \in \mathbf{R}^2$ is is given. Notice that in case $\varphi_{ij}(z,t) = \delta_{ij}\delta(z)$ then the previous invariance on the feature becomes the brightness invariance condition

$$\partial_t C_i + v_j \partial_j C_i = 0, \tag{14}$$

that is typically used to estimate the optical flow [28]. Here, the unknown is in fact the velocity field, whereas in the feature motion invariance condition (13) the unknown are the filters. This can be promptly seen when replacing Φ_i as in Eq. 3.1–(3) we get

$$\int_\Sigma \left(\partial_t \varphi_{ij} C_j + \varphi_{ij}\partial_t C_j + \varphi_{ij} v_k \partial_k C_j\right) dy = 0, \tag{15}$$

which holds for any $i = 1, \dots, n$ and $(x,t) \in \Pi$. Notice that this constraint is linear in the field φ. This can be interpreted by stating that learning under motion invariance, for any (x,t), consists of determining elements of the kernel of function

$$\mathscr{M}_{(x,t)}(\varphi_{ij}) := \int_\Sigma \left(\partial_t \varphi_{ij} C_j + \varphi_{ij}\partial_t C_j + \varphi_{ij} v_k \partial_k C_j\right) dy. \tag{16}$$

As we can promptly see $\mathscr{M}_{(x,t)}(\cdot)$ is defined by the knowledge of the video signal C and the by availability of the optical flow v. Depending on the color field C it quite easy to realize that $\mathscr{M}_{(x,t)}(\varphi_{ij})$ might be the null space, since while the possible visual configurations increase exponentially with the growth of the measure of Σ , the information associated with φ_{ij} only grows linearly the distance to the focus point. Hence condition (16) can be better satisfied in case of video with smooth spatiotemporal transitions. This is what happens for newborns, who experiment similar smooth transitions in early stage of development [11]. Moreover, sparseness of φ_{ij} also favors the satisfaction of (16). In particular, as will be better discussed in the remainder of the section, the satisfaction of motion invariance is favored by the receptive-field assumption. It is worth mentioning that the above constraints can be enforced at least in two different ways:

i. As stated above, we can impose constraint (16) for all points $(x,t) \in \Pi$. In doing so, one enforces motion invariance in any point of the retina;
ii. We can impose constraint (12) with the choice $x(t) = a(t)$, namely on the focus of attention trajectory.

In the reminder of this section we will follow the first approach, the second approach will be pursued in Section 3.3.1.

3.1.2. Analysis on the Continuous Retina

In the previous section we have discussed principles that drive the discovery of the filters φ_{ij} based on the MaxEnt principle and regularization. We provide

a soft-interpretation of the constraints, so as the adoption of these principles corresponds with the minimization of a functional that, following [5], it referred to as the "cognitive action":

$$\begin{aligned}\mathcal{A}_0(\varphi) = &\int_\Pi \Phi_i(\varphi)\,d\mu \log \int_\Pi \Phi_i(\varphi)\,d\mu - \lambda_C \int_\Pi \Phi_i(\varphi)\log\Phi_i(\varphi)\,d\mu \\ &+ \lambda_1 \int_\Pi \Big(\sum_{i=1}^{n} \Phi_i(\varphi) - 1\Big)^2 d\mu - \lambda_0 \int_\Pi \Phi_i(\varphi)[\Phi_i(\varphi) < 0]\,d\mu \\ &+ \frac{\lambda_P}{2}\int_\Pi (P_x\varphi_{ij}(x,t))^2\,\varpi(t)dtdx + \frac{\lambda_K}{2}\int_\Pi (P_t\varphi_{ij}(x,t))^2\,\varpi(t)dtdx \\ &+ \lambda_M \int_\Pi \big(\partial_t\Phi_i(\varphi) + v_j\partial_j\Phi_i(\varphi)\big)^2\,d\mu,\end{aligned} \tag{1}$$

where the notation $\Phi_i(\varphi)$ is used to stress the fact that Φ_i depends functionally on the filters φ. Here, if $\lambda_C = 1$, the first line is the negative of the mutual information and the constants $\lambda_C, \lambda_1, \lambda_0, \lambda_P, \lambda_K$, and λ_M are positive multipliers. This cognitive action can be given two different interpretations. First, one could think of the regularization terms and on the motion terms as penalty constraints, so as learning is interpreted in the classic framework of the MaxEnt principle. Second, we can (preferably) think of enriching the entropy with the regularization terms in the objective functions and regard motion term as the only actual constraint. Furthermore, notice that the mutual information (the first line) is rather involved, and it becomes too cumbersome to be used with a principle of least action. However, if we give up to attach the information-based terms their interpretation in terms of bits (or nats), we can rewrite the entropies that define the mutual information as

$$S(Y \mid X,T,F) \to -\int_D \Phi_i^2\,d\mu \quad \text{and} \quad S(Y) \to -\Big(\int_D \Phi_i\,d\mu\Big)^2. \tag{2}$$

Interestingly, this replacement does retain all the basic properties on the stationary points of the mutual information and, at the same time, it simplifies dramatically the overall action, which becomes

$$\begin{aligned}\mathcal{A}(\varphi) = &\frac{1}{2}\Big(\int_\Pi \Phi_i(\varphi)\,d\mu\Big)^2 - \frac{\lambda_C}{2}\int_\Pi \Phi_i^2(\varphi)\,d\mu \\ &+ \frac{\lambda_1}{2}\int_\Pi \Big(\sum_{i=1}^{n}\Phi_i(\varphi) - 1\Big)^2 d\mu - \lambda_0\int_\Pi \Phi_i(\varphi)\cdot[\Phi_i(\varphi) < 0]\,d\mu \\ &+ \frac{\lambda_P}{2}\int_\Pi (P_x\varphi_{ij})^2\,\varpi(t)dtdx + \frac{\lambda_K}{2}\int_\Pi (P_t\varphi_{ij})^2\,\varpi(t)dtdx \\ &+ \frac{\lambda_M}{2}\int_\Pi \big(\partial_t\Phi_i(\varphi) + v_j\partial_j\Phi_i(\varphi)\big)^2\,d\mu.\end{aligned} \tag{3}$$

$$\Phi_i(x,t) = \sigma(\varphi_{kj} * C_j)(x,t) = \sigma\Big(\int_\Sigma \varphi_{kj}(x-y,t)C_j(y,t)\,dy\Big)$$

In the following analysis we will consider the case in which $\sigma(\cdot)$ is the identity function, but the extension to the general case is straightforward. In order to be sure to preserve the commutativity of convolution — a property that in general holds when the integrals are extended to the entire plane — we have to make assumptions on the retina and on the domain on which the filters are defined. First of all assume that $\Pi = \Sigma_R \times [0,T]$, with $\Sigma_R = [-R,R]\times[-R,R]$, $R>0$; we will assume that C_i has spatial support in Σ_R and it is identically null outside, while φ_{ij} will be taken with spatial support in Σ_r with $0 < r \leq R$ and zero outside Σ_r. Under these assumption we can guarantee that the convolution $\varphi_{ij} * C_j$ is commutative in Σ_R. In particular, for all $x \in \Sigma_R$ we have

$$\begin{aligned}\Phi_i(x,t) = (\varphi_{ij} * C_j)(x,t) &= \int_{\Sigma_R} \varphi_{ij}(x-y,t) C_j(y,t)\, dy \\ &= \int_{\Sigma_R} \varphi_{ij}(y,t) C_j(x-y,t)\, dy = \int_{\Sigma_r} \varphi_{ij}(y,t) C_j(x-y,t)\, dy \\ &= (C_j * \varphi_{ij})(x,t).\end{aligned} \tag{4}$$

Before studying the stationarity of $\mathcal{A}$ we can conveniently elaborate its functional structure so as to get a more direct expression in terms of φ_{ij}. In particular, in order to provide an explicit expression of the motion term we need to introduce a number of coefficients that can be computed whenever we are given the video signal and the optical flow. Let us define*

$$\begin{aligned} W_{ml}(\xi,\zeta,\tau) &= f(z,\tau) C_m(z-\xi,\tau) C_l(z-\zeta,\tau) \\ Y_{ml}(\xi,\zeta,\tau) &= f(z,\tau)[\partial_\tau C_m(z-\xi,\tau) + v_\alpha \partial_\alpha C_m(z-\xi,\tau)] C_l(z-\zeta,\tau) \\ H_{ml}(\xi,\zeta,\tau) &= f(z,\tau)[\partial_\tau C_m(z-\xi,\tau) + v_\alpha \partial_\alpha C_m(z-\xi,\tau)]\cdot \\ &\qquad \cdot [\partial_\tau C_l(z-\zeta,\tau) + v_\beta \partial_\beta C_l(z-\zeta,\tau)]. \end{aligned} \tag{5}$$

In case of still images we can promptly see that only $W_{ml}(\xi,\zeta,\tau) \neq 0$. Its value turns out to be a sort of autocorrelation of the color field, which operates over the different channels m,l, as well as at spatial level between the values at ξ and ζ. The coefficients $Y_{ml}(\xi,\zeta,\tau)$, $H_{ml}(\xi,\zeta,\tau)$ are affected by motion but have a related autocorrelation meaning. Once, we introduce these coefficients, the following property can be stated.

Proposition A. *Motion term $\omega(\varphi)$ turns out to be a quadratic function of φ and its temporal derivative, that is (integral over τ is understood)*

$$\begin{aligned}\omega(\varphi) = \frac{1}{2}\big(&\partial_\tau \varphi_{km}(\xi,\tau) W_{ml}(\xi,\zeta,\tau) \partial_\tau \varphi_{kl}(\zeta,\tau) \\ &+ 2\varphi_{km}(\xi,\tau) Y_{ml}(\xi,\zeta,\tau) \partial_\tau \varphi_{kl}(\zeta,\tau) + \varphi_{km}(\xi,\tau) H_{ml}(\xi,\zeta,\tau) \varphi_{kl}(\zeta,\tau)\big).\end{aligned} \tag{6}$$

Proof. The proof arises from plugging expression of the features into the motion term. ∎

* We use Einstein's notation also on continuum spatial indexes á la Weinberg (see [59]).

The statement of the Euler-Lagrange equations also benefits from defining

$$\begin{aligned}\Xi_{jk}(x,\xi,t) &= -W_{jk}(\xi,x,t)\\ \Theta_{jk}(x,\xi,t) &= Y_{jk}(x,\xi,t) - Y_{kj}(\xi,x,t) - \partial_t W_{kj}(\xi,x,t)\\ \Upsilon_{jk}(x,\xi,t) &= H_{kj}(\xi,x,t) - \partial_t Y_{kj}(\xi,x,t).\end{aligned} \tag{7}$$

In addition, based on $\Xi_{jk}(x,\xi,t), \Theta_{jk}(x,\xi,t)$, and $\Upsilon_{jk}(x,\xi,t)$, we also introduce (no integral over t)

$$\begin{aligned} c_j(x,t) &:= f(z,t)C_j(z-x,t),\\ T_{jk,im}(x,\xi,t,\partial_t) &:= \lambda_C \Xi_{jk}(x,\xi,t)\delta_{im} + \sum_{l=1}^{n} \Xi_{jk}(x,\xi,t)\delta_{\ell m}\\ &\quad + \lambda_M(\Xi_{jk}(x,\xi,t)\partial_t^2 + \Theta_{jk}(x,\xi,t)\partial_t + \Upsilon_{jk}(x,\xi,t))\delta_{im},\\ \Delta_{jk,im}(x,\xi,t,\partial_t) &:= T_{jk,im}(x,\xi,t,\partial_t) + c_j(x,t)c_k(\xi,t)\delta_{im},\\ \rho_{ij}(x,t) &:= -\lambda_1 c_j(x,t) - \lambda_0 f(z,t)C_j(z-x,t)[\Phi_i(z,t)<0]. \end{aligned} \tag{8}$$

In what follows we will regard ρ as a function that is independent of the variables* φ. We are now ready to express the stationary condition of the action (3).

Theorem A. *The stationarity conditions of* (3) *leads to the following Euler-Lagrange equations in the filters* φ_{ij}

$$\begin{aligned}&\lambda_K P_t^*(\varpi(t)P_t\varphi_{ij}(x,t)) + \lambda_P \varpi(t) P_x^* P_x \varphi_{ij}(x,t) + T_{jk,im}(x,\xi,t,\partial_t)\varphi_{mk}(\xi,t)\\ &\quad + c_j(x,t)c_k(\xi,\tau)\varphi_{ik}(\xi,\tau) + \rho_{ij}(x,t) = 0,\end{aligned} \tag{9}$$

where T *and* ρ *are defined in Eq.* (8).

Proof. The Euler-Lagrange equation of the action arises from $\delta\mathcal{A}(\varphi)/\delta\varphi_{ij}(x,t) = 0$. So we need to take the variational derivative of all the terms of action in Eq. (3). In the following calculation, we will assume that $d\mu(x,t) = f(x,t)\,dx\,dt$. The first term yields

$$\begin{aligned}\Big(\int_\Pi \Phi_k\, d\mu\Big)\frac{\delta}{\delta\varphi_{ij}(x,t)} f(z,\tau)\varphi_{km}(y,\tau)C_m(z-y,\tau)\\ = f(y,t)C_j(y-x,t)f(z,\tau)\varphi_{ik}(\xi,\tau)C_k(z-\xi,\tau);\end{aligned} \tag{10}$$

while the second term gives

$$\begin{aligned}\frac{\delta}{\delta\varphi_{ij}(x,t)}\frac{1}{2}\int_\Pi \Phi_k^2\, d\mu &= f(z,t)\Phi_i(z,t)C_j(z-x,t)\\ &= f(z,t)C_j(z-x,t)C_k(z-\xi,t)\varphi_{ik}(\xi,t).\end{aligned} \tag{11}$$

* Actually ρ depends on φ through the step function $[\Phi \le 0]$, so that the precise statement would be that ρ is independent of φ in the regions with definite sign of the feature φ. This can be avoided if we impose the perfect satisfaction of the normalization conditions or if we assume a softmax normalization of the features., this is also the reason why we do not bother too much for analyzing more carefully the regularity properties of φ

The variation of the third term similarly yields

$$\sum_{m=1}^{n} f(z,t)C_j(z-x,t)C_k(z-\xi,t)\varphi_{mk}(\xi,t) - f(z,t)C_j(z-x,t). \tag{12}$$

The variation of the terms that implements positivity is somewhat more involved:

$$\begin{aligned}\frac{\delta}{\delta\varphi_{ij}(x,t)}\int_\Pi \Phi_k[\Phi_k<0]\,d\mu =& \frac{\delta\Phi_k(z,\tau)}{\delta\varphi_{ij}(x,t)}[\Phi_k(z,\tau)<0]f(z,\tau)\\ &+\Phi_k(z,\tau)\frac{\delta[\Phi_k(z,\tau)<0]}{\delta\varphi_{ij}(x,t)}f(z,\tau)\,dz\,d\tau.\end{aligned} \tag{13}$$

However, the second term is zero since

$$\begin{aligned}\Phi_k(z,\tau)\delta[\Phi_k(z,\tau)<0]f(z,\tau) =& \varphi_{km}(\xi,\tau)C(z-\xi,\tau)\\ &\cdot\Big(\big[\varphi_{km}(y,\tau)C_m(z-y,\tau)\\ &\quad+\epsilon\delta\varphi_{km}(y,\tau)C_m(z-y,\tau)<0\big]\\ &\quad-\big[\varphi_{km}(y,\tau)C_m(z-y,\tau)<0\big]\Big).\end{aligned} \tag{14}$$

The difference of the two Iverson's brakets is always zero unless the epsilon-term makes the argument of the first braket have an opposite sign with respect to the second. Since ϵ is arbitrary small, this can only happen if $\varphi_{km}(\xi,\tau)C_m(z-\xi,\tau) = 0$. Thus in either cases the whole term vanishes. Hence, we get

$$\frac{\delta}{\delta\varphi_{ij}(x,t)}\int_\Pi \Phi_k\cdot[\Phi_k<0]\,d\mu = f(z,t)C_j(z-x,t)[\Phi_i(z,t)<0]. \tag{15}$$

Recalling (6) and definition (7) we can eventually compute the variation of the motion invariance term. Since we can always assume that at $t=0$ and $t=T$ that the video C with its derivative is identically zero, we automatically have $W_{ml}(\xi,x,0)\equiv W_{ml}(\xi,x,T)\equiv Y_{ml}(\xi,x,0)\equiv Y_{ml}(\xi,x,T)\equiv H_{ml}(\xi,x,0)\equiv H_{ml}(\xi,x,T)\equiv 0$. This properties save us from having boundary terms coming from the integration by parts that we need to perform when we compute the variation of this term.

$$\begin{aligned}\frac{\delta\omega(\varphi)}{\delta\varphi_{ij}(x,t)} =& -W_{lj}(\xi,x,t)\partial_t^2\varphi_{il}(\xi,t)\\ &+(Y_{jl}(x,\xi,t)-Y_{lj}(\xi,x,t)-\partial_t W_{lj}(\xi,x,t))\partial_t\varphi_{il}(\xi,t)\\ &+(H_{lj}(\xi,x,t)-\partial_t Y_{lj}(\xi,x,t))\varphi_{il}(\xi,t).\end{aligned} \tag{16}$$

Hence

$$\frac{\delta\omega(\varphi)}{\delta\varphi_{ij}(x,t)} = \big(\Xi_{jk}(x,\xi,t)\partial_t^2+\Theta_{jk}(x,\xi,t)\partial_t+\Upsilon_{jk}(x,\xi,t)\big)\varphi_{ik}(\xi,t). \tag{17}$$

In these calculations we have used intensively the commutative property of the convolution as stated in Eq. (4), which allows us to avoid expressions with an higher degree of space non-locality. Then the Euler-Lagrange equations reads:

$$\begin{aligned}
&\lambda_K P_t^*(\varpi(t)P_t\varphi_{ij}(x,t)) + \lambda_P \varpi(t) P_x^* P_x \varphi_{ij}(x,t) \\
&+ c_j(x,t)\big(c_k(\xi,\tau)\varphi_{ik}(\xi,\tau) - \lambda_1\big) + \lambda_C \Xi_{jk}(x,\xi,t)\varphi_{ik}(\xi,t) \\
&\quad - \lambda_1 \sum_{m=1}^{n} \Xi_{jk}(x,\xi,t)\varphi_{mk}(\xi,t) - \lambda_0 f(z,t) C_j(z-x,t)[\Phi_i(z,t) < 0] \\
&\quad + \lambda_M \big(\Xi_{jk}(x,\xi,t)\partial_t^2 + \Theta_{jk}(x,\xi,t)\partial_t + \Upsilon_{jk}(x,\xi,t)\big)\varphi_{ik}(\xi,t) = 0,
\end{aligned} \tag{18}$$

which is nothing else than Eq. (9). ∎

Boundary conditions. In order to be solved, E-L equations (9) require the definition of the boundary conditions on Π. Clearly the mutual information term does not add any boundary conditions to the E-L equations and we have already discussed that we can choose C and its derivative to be 0 at $t = T$ so that the motion term does not add any conditions on the boundaries either. As we will see in details in the following sections and as we have already discussed extensively in Chapter 2, however, boundary conditions will appear that are due to the temporal regularization term.

Non-locality and ill-position. This theorem shows that the EL-equations are non-local integro-differential equations. Notice that Eq. (9) is non-local in both spatial (third and fourth terms) and time (fourth term). This result suggests that an agent designed on the basis of Eq. (9) would be doomed to fail, since its solution is inherent intractable in terms of computational complexity. Basically, the lack of locality, makes Eq. (9) unsuitable to model the emergence of visual features in nature. In what follows we will show how to overcome this critical complexity issues by modifying the position of the problem of visual feature so as to make it well-posed.

Temporal locality. Other than the non causality due to the presence of boundary conditions in the theory that we have described so far there is a much more explicit sorce of non-causality due to the temporal non-locality of the lagrangian that we are using. From Eq. (9) we immediately see that the last term is non-local in time; this, as a consequence, means that the equations are non-causal. This is basically due to the need of knowing the probability of the hidden symbols to determine the entropy. Formally, the probability of the symbols does require to know all the video over the life interval $[0, T]$, which breaks temporal locality. This problem can be faced in different ways:

i. Enforce time locality by computing the entropy by splitting the averaging on frames and time as follows:

$$S(Y) \to \int_0^T \left(\int_\Sigma g(x - a(t))\Phi_i(x,t) f(x,t)\, dx \right)^2 \varpi(t) dt. \tag{19}$$

Clearly this way of splitting the measure $d\mu$ only approximates the actual entropy of the source. When averaging at frame level one might get a biased

view on the probability of the symbols that, however, is somewhat balanced by the temporal average over all the time horizon.

ii. Define the following estimation of the probability of symbol i at t:

$$s_i(t) = \int_0^t \int_\Sigma \Phi_i(x,\tau) f(x,\tau)\, dx d\tau = \int_0^t \int_\Sigma \Phi_i(x,\tau) \varpi(\tau) g(x - a(\tau))\, dx d\tau. \tag{20}$$

and express the entropy on the basis of this estimation instead of the actual value of the probability of symbol i given by $\int_\Pi d\mu \Phi_i$. In this way the entropy term $S(Y)$ in the Lagrangian can be replaced with

$$\tilde{S}(Y) = \frac{1}{T} \int_0^T s_i^2(t)\, dt + \alpha \int_0^T \left(s_i(t) - \int_0^t \int_X \Phi_i(x,\tau) f(x,\tau)\, dx d\tau \right)^2 dt \tag{21}$$

where the second term, with an appropriate non-negative α is required to enforce the constraint on the value gained by $s_i(t)$.

iii. Let us consider the above causal entropy $\tilde{S}_Y$ given by Eq. (21) and enforce a differential form of the the constraint on $s_i(t)$. In doing so, the entropy term in the Lagrangian can be replaced with

$$\tilde{\tilde{S}}(Y) = \frac{1}{T} \int_0^T s_i^2(t)\, dt + \alpha \int_0^T \left(\dot{s}_i(t) - \int_\Sigma \Phi_i(x,t) f(x,t)\, dx \right)^2 dt. \tag{22}$$

Clearly, in doing so, unlike the formulation based on the cognitive action (3), the corresponding E-L equations that we derive are local in time. However, we need to involve the auxiliary variable s_i in addition to the other Lagrangian coordinates.

Interestingly, $\tilde{S}_Y$ offers a consistent asymptotic approximation of $S(Y)$. In particular, the following results connects the two terms.

Proposition B. *If* $\lim_{t\to\infty} s_i(t) = p_i(T) := \int_\Sigma \int_0^T \Phi_i(x,t) f(x,t)\, dx dt$, *then*

$$\lim_{T\to+\infty} \left| p_i^2(T) - \frac{1}{T} \int_0^T s_i^2(t)\, dt \right| = 0. \tag{23}$$

Proof. From the hypothesis $\forall \epsilon > 0$ there exists T_ϵ such that $\forall t > T_\epsilon : \ |p_i - s_i(t)| \leq \epsilon$

$$\begin{aligned}
\alpha_S(T) &= \frac{1}{T} \left| T p_i^2 - \int_0^T s_i^2\, dt \right| = \frac{1}{T} \left| \int_0^T p_i^2\, dt - \int_0^T s_i^2\, dt \right| \\
&= \frac{1}{T} \left| \int_0^T (p_i + s_i)(p_i - s_i)\, dt \right| \leq \frac{1}{T} \int_0^T (p_i + s_i)|p_i - s_i|\, dt \\
&\leq \frac{1}{T} \int_0^{T_\epsilon} (p_i + s_i)|p_i - s_i|\, dt + \frac{1}{T} \int_{T_\epsilon}^T (2p_i + \epsilon)\epsilon\, dt \\
&\leq 2\frac{T_\epsilon}{T}\epsilon + \frac{T - T_\epsilon}{T}(2 + \epsilon)\epsilon < \left(2\frac{T_\epsilon}{T} + (2 + \epsilon) \right)\epsilon.
\end{aligned}$$

Now, for any $\delta > 0$ the condition $\alpha_S(T) < \delta$ yields $2\frac{T_\epsilon}{T} + (2+\epsilon)\epsilon < \delta$ which is satisfied when choosing

$$\epsilon < \sqrt{\left(1 + \frac{T_\epsilon}{T}\right)^2 + \delta} - \left(1 + \frac{T_\epsilon}{T}\right)$$

and $T > T_\epsilon$. ∎

We are now ready to see how how the Euler-Lagrange equations are transformed once time-locality is handled. In particular, in the following, we consider the case i, but extension to ii. and iii. can be considered.

Theorem B. *The functional $\mathcal{A}(\varphi)$ under the replacement described in Eq.* (19) *admits time-local E-L equations, i.e Eq.* (9) *becomes*

$$\lambda_K P_t^*(\varpi(t) P_t \varphi_{ij}(x,t)) + \lambda_P \varpi(t) P_x^* P_x \varphi_{ij}(x,t) + \Delta_{jk,im}(x,\xi,t,\partial_t)\varphi_{mk}(\xi,t) + \rho_{ij}(x,t) = 0, \tag{24}$$

Proof. It is sufficient to replace the variation of the energy term, which is now dramatically simplified (no integral over t)

$$\frac{\delta}{\delta\varphi_{ij}(x,t)} \int_0^T \left(\int_X \Phi_i(z,\tau) f(z,\tau)\, dz \right)^2 d\tau = c_j(x,t) c_k(\xi,t) \varphi_{ik}(\xi,t).$$

Finally, the theorem arises when considering the definitions (8). ∎

It is easy to see that temporal locality can also be gained in the case in which the entropy is defined according to Eq. (22).

Space locality. We will now show how to gain space locality, which is still missing in Eq. (24). The intuition is that the lack of space locality is inherently connected with the definition of convolutional features, whenever one makes no delimitation on the context required to compute the features. As already pointed when addressing motion invariance, while the possible visual configurations increase exponentially with the growth of the measure of Σ , the information associated with φ_{ij} only grows linearly the distance to the focus point. We will make use of a generalized notion of *receptive field* that, as it will be proven in the following, allows us to gain spatial locality.

To be more precise assume the following factorization for the filters

$$\varphi_{ij}(x,t) = G(x)\phi_{ij}(x,t), \tag{25}$$

where $G\colon \Sigma \to \mathbf{R}$ is a smooth, bell shaped function. Notice that this corresponds with expressing the computation of the features by

$$\Phi_i(x,t) = \int_\Sigma G(y)\phi_{ij}(y,t) C(x-y,t)\, dy. \tag{26}$$

In so doing, the contribution of the color field at distance $x-y$ is weighed on the basis of the receptive field structure induced by bell-shaped function G. Then the non-local term in Eq. (24) reads $G(\xi)\Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t)$.

Theorem C. *Let $G\colon \Sigma \to \mathbf{R}$ be the Green function of an self-adjoint operator L and let $G(\partial\Sigma) = 0$, where $\partial\Sigma$ denotes the boundary of Σ. Then Eq. (24) is equivalent to the following (local) system of differential equations:*

$$\begin{cases} \lambda_K P_t^*(\varpi(t)P_t G(x)\phi_{ij}(x,t)) + \lambda_P \varpi(t) P_x^* P_x G(x)\phi_{ij}(x,t) + \Lambda_{ij}(x,0,t) + \rho_{ij}(x,t) = 0; \\ L\Lambda_{ij}(x,\xi,t) = \Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t). \end{cases} \tag{27}$$

Proof. Let $\Lambda_{ij}(x,\xi,t)$ be a solution of the differential equation

$$L\Lambda_{ij}(x,\xi,t) = \Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t),$$

where L is a self-adjoint operator. The term $G(\xi)\Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t)$ becomes (no integral over t)

$$G(\xi)\Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t) = G(\xi)L\Lambda_{ij}(x,\xi,t). \tag{28}$$

Now, since L is self-adjoint, we have $L^*G = LG = \delta$ and, consequently, we get (no integral over t)

$$G(\xi)\Delta_{jk,im}(x,\xi,t,\partial_t)\phi_{mk}(\xi,t) = \Lambda_{ij}(x,0,t), \tag{29}$$

which is a local expression in space. Finally, Eq. (24) turns out to be equivalent to Eq. (27). ∎

These differential equations, along with their boundary conditions, can be thought of as information-based laws that dictate the spatiotemporal behavior of the visual filters. Notice that space locality has been gained at the price of enriching the space by the adjoint variable Λ_{ij}. It contributes to face and break chicken-egg dilemma on whether we first need to define the context for computing the related visual feature or if the feature does in fact define also the context from which it is generated. The transformation of Eq. (24) (integro-differential equations) into Eq. (27) (differential equations) is paid by the introducing of the cyclic computational structure of Eq. (27) that, however, is affordable from a computational point of view. It is worth mentioning that from an epistemological point of view, Eq. (27) comes from variational principles that very much remind us the scheme used in physics; for this reason we use the term information-based laws of visual features. Clearly, we can always read these differential equations as a computational model of learning visual features.

The following theorem gives insights on the possibility of finding G and L that satisfy the properties required by Theorem C with arbitrary precision.

Theorem D. *Let $G_\sigma(x)$ be a gaussian with variance σ and zero mean; et $L_\sigma^m := \sum_{n=0}^m (-1)^n(\sigma^{2n}/2^n n!)\nabla^{2n}$, then G_σ and L_σ^m satisfy the hypothesis of Theorem C if σ is chosen small enough. More precisely we have that*

$$\lim_{\sigma\to 0} \int \big(L_\sigma^m G_\sigma(x)\big)\varphi(x)\,dx = \varphi(0), \qquad \forall \varphi \in C_c^\infty(\mathbf{R}). \tag{30}$$

Proof. See Appendix B

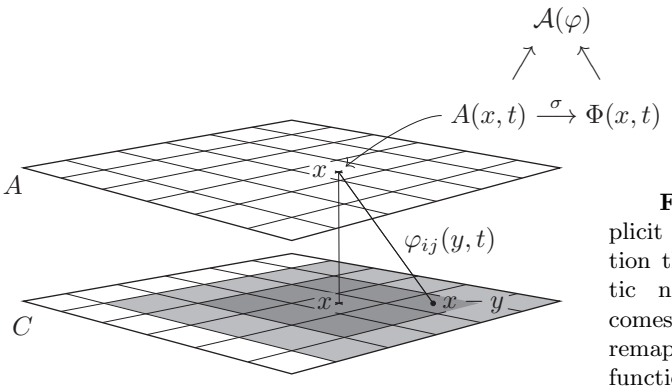


Fig. 12. Without explicit constraints in the action that enforces probabilistic normalization the outcomes of convolution must be remapped with a nonlinear function σ (for example a softmax function) in order to be used to built the information based part of the functional index.

This result expressed by this theorem makes the reduction of Eq. (28) possible in case we adopt *receptive fields*. Let $\rho_\sigma(x) := L_\sigma^m G_\sigma(x)$ be. In Appendix A we can see that, for a given m we have that $\rho_\sigma(x)$ approaches the δ distribution as $\sigma \to 0$. Basically, we meet the assumption of Theorem C for finite m, which is a crucial computational issue concerning the adjoint equation $L\Lambda_{ij}(x,\xi,t) = \Delta_{jq,ip}(x,\xi,t,\partial_t)\phi_{pq}(\xi,t)$. As stated by the theorem, this holds for "small" σ, that can be regarded as a receptive field assumption.

It is interesting to notice that the property claimed in the theorem works also if G is not itself a Green's function but in case it is a linear combination of Green's functions evaluated at different points, that is

$$G(x) = \sum_{i=1}^{N} \alpha_i J(x - x_i), \tag{31}$$

so as Eq. (25) is in fact quite general in terms of function representation. However, it is evident that as N increases also the number of terms in Eq. (27) does the same, so that it might indicate that the resolution of such equations becomes harder.

Softmax formulation and focus of attention. Instead of imposing probabilistic normalization implicitly, we can express the constraints by classic softmax as follows:

$$A_i(x,t) := \int_\Sigma \varphi_{ij}(y,t) C_j(x-y,t)\, dy, \qquad \Phi_i(x,t) = \sigma_i(A_1(x,t),\ldots,A_n(x,t)), \tag{32}$$

where $\sigma_i(x_1,\ldots,x_n) := e^{x_i} / \sum_{k=1}^n e^{x_k}$. With this redefinition, the the information theory based terms of the action are automatically well-defined, while the motion invariance term can still be imposed on the convolutional activations

$A_i(x,t)$. This formulation therefore it is based on the following action

$$\begin{aligned}\mathcal{A}(\varphi) =& \frac{1}{2}\Big(\int_\Pi \Phi_i \, d\mu\Big)^2 - \frac{\lambda_C}{2}\int_\Pi \Phi_i^2 \, d\mu \\ &+ \frac{\lambda_P}{2}\int_\Pi (P_x\varphi_{ij}(x,t))^2 \varpi(t)\, dtdx + \frac{\lambda_K}{2}\int_\Pi (P_t\varphi_{ij}(x,t))^2 \varpi(t)\, dtdx \\ &+ \frac{\lambda_M}{2}\int_\Pi \big(\partial_t A_i(\varphi) + v_j \partial_j A_i(\varphi)\big)^2 \, d\mu.\end{aligned} \tag{33}$$

that gives rise to EL-equations very related to Eq. (27).

3.2. NEURAL INTERPRETATION ON THE RETINA

So far, we have developed a field theory on Π, now we will reformulate it on a discretized retina $\Sigma^\sharp$. We need to see how this fields can be re-written on a discretized retina $\Sigma^\sharp = \{(i,j) \mid 0 \le i < \ell, \quad 0 \le j < \ell\}$. While the filters are characterized by $\varphi_{ijx}(t)$, the color field will be replaced with $C_{ix}(t)$.

Notice that, because of the factorization $f(x,t) = \varpi(t)g(x-a(t))$, the term g_x in the discretized formulation is also a function of time, which will turn out to contribute to the time dependence that affects the coefficients of the differential equation that governs the evolution of the filters. Moreover, since g_x plays the role of a probability distribution over the retina, for every t, we have $\sum_{x\in\Sigma^\sharp} g_x = 1$. As a consequence this yields $\int_0^T \varpi(t) = 1$.

On the discrete retina the activations as defined in Eq. 3.1.2–(32) assume the form (we do not explicitly write the time dependence on the time)

$$A_{ix_1x_2} = \varphi_{ij\xi_1\xi_2} C_{j(x_1-\xi_1)(x_2-\xi_2)}. \tag{1}$$

Now let $\gamma^x := (C_{1(x_1-1)(x_2-1)}, C_{1(x_1-1)(x_2-2)}, \dots, C_{m(x_1-\ell)(x_2-\ell)}) \in \mathbf{R}^{m\ell^2}$ and $\chi^i := (\varphi_{i111}, \varphi_{i112}, \dots, \varphi_{im\ell\ell}) \in \mathbf{R}^{m\ell^2}$. Then

$$A_{ix} \equiv A_{ix_1x_2} = \chi^i_\alpha \gamma^x_\alpha. \tag{2}$$

Let us now see how the Cognitive Action can be written in terms of these discretized variables. First of all let us analyze the motion-invariance term. If we let ζ^x to be the vector that for each pixel x on the retina collects the components of the discretization of the term $v(x,t)\cdot\nabla_x C_j(x-\xi,t)$ with respect to the indexes ξ and j, then the part of the Lagrangian relative to the motion invariance term can be written as

$$\frac{1}{2}\int_0^T \varpi(t) g_x \Big((\chi^i_\alpha \gamma^x_\alpha)^\cdot + \chi^i_\alpha \zeta^x_\alpha \Big)^2. \tag{3}$$

The square in the previous equation, once expanded, gives:

$$\begin{aligned} g_x \Big((\chi^i_\alpha \gamma^x_\alpha)^\cdot + \chi^i_\alpha \zeta^x_\alpha \Big)^2 =& \chi^i_\alpha \big(g_x(\dot\gamma^x_\alpha \dot\gamma^x_\beta + \zeta^x_\alpha \zeta^x_\beta + 2\dot\gamma^x_\alpha \zeta^x_\beta) \big) \delta_{ij} \chi^j_\beta \\ &+ 2\chi^i_\alpha \big(g_x(\dot\gamma^x_\alpha \gamma^x_\beta + \zeta^x_\alpha \gamma^x_\beta) \big) \delta_{ij} \dot\chi^j_\beta \\ &+ \dot\chi^i_\alpha \big(g_x \gamma^x_\alpha \gamma^x_\beta \big) \delta_{ij} \dot\chi^j_\beta \\ =& \chi^i_\alpha O_{\alpha\beta} \delta_{ij} \chi^j_\beta + 2\chi^i_\alpha N_{\alpha\beta} \delta_{ij} \dot\chi^j_\beta + \dot\chi^i_\alpha M_{\alpha\beta} \delta_{ij} \dot\chi^j_\beta. \end{aligned} \tag{4}$$

Where we have defined $O_{\alpha\beta} := g_x(\dot\gamma^x_\alpha \dot\gamma^x_\beta + \zeta^x_\alpha \zeta^x_\beta + 2\dot\gamma^x_\alpha \zeta^x_\beta)$, $N_{\alpha\beta} := g_x(\dot\gamma^x_\alpha \gamma^x_\beta + \zeta^x_\alpha \gamma^x_\beta)$ and $M_{\alpha\beta} := g_x \gamma^x_\alpha \gamma^x_\beta$.

Given $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{n\times k}$ and having defined the vectorization operation as follows

$$\operatorname{vec}(A) = (a_{11}, a_{21}, \dots, a_{m1}, a_{m2}, a_{12}, a_{22}, \dots, a_{mn})', \tag{5}$$

this two identities holds

1. $\operatorname{vec}(AB) = (B' \otimes \operatorname{Id}_m) \operatorname{vec}(A)$;

2. $\mathrm{Tr}(A'B) = \mathrm{vec}(A) \cdot \mathrm{vec}(B)$,

where $\otimes$ is the Kronecker product. Using 1. and 2. we can rewrite the terms in Eq. (4) as follows:

$$\begin{aligned}
&\chi^i_\alpha O_{\alpha\beta}\delta_{ij}\chi^j_\beta = \mathrm{Tr}(\chi'\chi O) = \mathrm{vec}(\chi)\cdot\mathrm{vec}(\chi O) = \mathrm{vec}(\chi)\cdot(O'\otimes \mathrm{Id}_{m\ell^2})\,\mathrm{vec}(\chi);\\
&\chi^i_\alpha N_{\alpha\beta}\delta_{ij}\dot\chi^j_\beta = \mathrm{Tr}(\chi'\dot\chi N) = \mathrm{vec}(\chi)\cdot\mathrm{vec}(\dot\chi N) = \mathrm{vec}(\chi)\cdot(N'\otimes \mathrm{Id}_{m\ell^2})\,\mathrm{vec}(\dot\chi);\\
&\dot\chi^i_\alpha M_{\alpha\beta}\delta_{ij}\dot\chi^j_\beta = \mathrm{Tr}(\dot\chi'\dot\chi M) = \mathrm{vec}(\dot\chi)\cdot\mathrm{vec}(\dot\chi M) = \mathrm{vec}(\dot\chi)\cdot(M'\otimes \mathrm{Id}_{m\ell^2})\,\mathrm{vec}(\dot\chi).
\end{aligned} \tag{6}$$

Once we define $q := \mathrm{vec}(\chi)$, $O^\natural := (O'\otimes \mathrm{Id}_{m\ell^2})$, $N^\natural := (N'\otimes \mathrm{Id}_{m\ell^2})$ and $M^\natural := (M'\otimes \mathrm{Id}_{m\ell^2})$ we eventually have

$$\chi^i_\alpha O_{\alpha\beta}\delta_{ij}\chi^j_\beta = q\cdot O^\natural q, \quad \chi^i_\alpha N_{\alpha\beta}\delta_{ij}\dot\chi^j_\beta = q\cdot N^\natural \dot q, \quad \dot\chi^i_\alpha M_{\alpha\beta}\delta_{ij}\dot\chi^j_\beta = \dot q\cdot M^\natural \dot q. \tag{7}$$

Notice how these matrices are the discrete counterparts of the functions W, Y, H defined in the previous section. Furthermore we group the terms of the mutual information in the potential-like term $U(q,t)$ defined as follows:

$$U(q,t) = \frac{1}{2}\big(g_x\sigma_i(A_x)\big)^2 - \frac{\lambda_C}{2}g_x\Big(\sigma_i\big(A_x\big)\Big)^2. \tag{8}$$

The above derivations can be summed up in the following proposition

Proposition C. *On the discrete retina the functional*

$$\frac{1}{2}\Big(\int_\Pi \Phi_i\,d\mu\Big)^2 - \frac{\lambda_C}{2}\int_\Pi \Phi_i^2\,d\mu + \frac{\lambda_M}{2}\int_\Pi \big(\partial_t A_i(\varphi) + v_j\partial_j A_i(\varphi)\big)^2\,d\mu, \tag{9}$$

which is the Cognitive Action in 3.1.2–Eq. (33) *without the regularization terms, becomes*

$$\mathcal{V}(q) = \int_0^T \varpi(t)U\big(q,t\big)\,dt + \lambda_M\mathcal{M}(q) \tag{10}$$

where

$$\mathcal{M}(q) := \int_0^T \varpi(t)\left(\frac{1}{2}\dot q M^\natural(t)\dot q + qN^\natural(t)\dot q + \frac{1}{2}q(t)O^\natural(t)q(t)\right)\,dt. \tag{11}$$

We will now show that if we pair the functional (10) with the regularization term

$$\mathcal{R}(q) := \int_0^T \varpi(t)dt\,\left(\frac{\alpha}{2}|\ddot q(t)|^2 + \frac{\beta}{2}|\dot q(t)|^2 + \frac{1}{2}|\gamma_1\dot q(t) + \gamma_2\ddot q(t)|^2 + \frac{k}{2}|q|^2\right), \tag{12}$$

then the resulting cognitive action

$$\Gamma(q) := \mathcal{V}(q) + \mathcal{R}(q) \tag{13}$$

admits a minimum.

In order to understand the peculiar structure of the chosen regularization term notice that if we pose $\mu = \alpha+\gamma_2^2$, $\nu = \beta+\gamma_1^2$, $\gamma = \gamma_1\cdot\gamma_2$ then Eq. (13) can

be rewritten as

$$\Gamma(q) = \int_0^T \varpi\Big(\frac{\mu}{2}|\ddot{q}|^2 + \frac{\nu}{2}|\dot{q}|^2 + \gamma\dot{q}\cdot\ddot{q} + \frac{k}{2}|q|^2 + U\big(q, C\big)\Big)\, dt + \lambda_M \mathcal{M}(q). \quad (14)$$

As we have argue in Chapter 2 the interpretation of learning by means of functional (14) is especially interesting since, unlike the case of the classic action in mechanics, it admits a minimum under appropriate conditions.

The following theorem, that is a straightforward extension of a results discussed in Chapter 2, sections 2.1 and 2.2 and appeared in [7], offers an important result on the well-posedness of learning.

Theorem E. *If the following coercivity conditions**

$$\mu > \gamma_2^2, \quad \nu > \gamma_1^2, \quad k > 0 \quad (15)$$

hold true then functional Γ*, defined by Eq.* (14)*, admits a minimum on the set*

$$\mathbf{X} = \{\, q \in H^2((0,T), \mathbf{R}^n) \mid q(0) = q^0, \dot{q}(0) = q^1 \,\}. \quad (16)$$

Proof. The proof follows the one in [7] with the additional observation that $\mathcal{M}(q) \geq 0$ and that it contains at most the first derivatives of q. ∎

Euler-Lagrange Equations. For the porpuse of taking the variation of the functional Γ it is convenient to rearrange it so to have all the terms with at least one derivative all grouped together: $\Gamma(q) = \Gamma_1(q) + \Gamma_2(q)$ with

$$\Gamma_1(q) = \int_0^T \left(\hat{U}(q,t) + \frac{1}{2}q(t)(\hat{O}^\natural(t) + \hat{k})q(t)\right)\, dt, \quad (17)$$

and

$$\Gamma_2(q) := \int_0^T dt\, \Big(\frac{\hat{\mu}}{2}|\ddot{q}|^2 + \frac{\hat{\nu}}{2}|\dot{q}|^2 + \hat{\gamma}\dot{q}\cdot\ddot{q} + \frac{\lambda_M}{2}\dot{q}\cdot\hat{M}^\natural\dot{q} + \lambda_M q\cdot\hat{N}^\natural\dot{q}\Big). \quad (18)$$

We have also introduced the following notation: for any expression A we let $\hat{A}(t) := \varpi(t)A$. In what follows we will also assume

$$\varpi(t) = \frac{\theta}{e^{\theta T} - 1} e^{\theta t} \quad (19)$$

with $\theta > 0$. In general, $\varpi(t)$ needs to be monotone increasing, so as to yield dissipation. For the moment therefore we are disregarding the causality issues that we raised in Section 2.1 and that we are going to address again at the end of this section and in the next section as well. With this factorization we see that the variation of Γ_1, other than being immediate, does not give any extra boundary condition. So let us focus on the variation of $\Gamma_2(q)$.

Let us consider the variation v and define $\psi(s) = \Gamma_2(q + sv)$, where $s \in \mathbf{R}$. In the analysis below, we will repeatedly use the fact that $v(0) = \dot{v}(0) = 0$. This corresponds with the assignment of the initial values $q(0)$ and $\dot{q}(0)$. Since we want to provide a causal computational framework for $q(t)$, this is in fact the

* These conditions are indeed equivalent to $\alpha > 0$,$\beta > 0$ and $k > 0$.

first step towards this direction. The stationarity condition for the functional Γ_2 is $\psi'(0) = 0$,

$$\psi'(0) = \int_0^T dt \left\{(\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})\cdot\ddot{v}+[(\hat{\nu}+\lambda_M\hat{M}^\natural)\dot{q}+\hat{\gamma}\ddot{q}+\lambda_M(\hat{N}^\natural)'q]\cdot\dot{v}+\lambda_M\hat{N}^\natural\dot{q}\cdot v\right\} \quad (20)$$

With a few integration by parts we get

$$\begin{aligned}\psi'(0) =& \left[(\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})\dot{v} + \left((\hat{\nu}+\lambda_M\hat{M}^\natural)\dot{q}+\hat{\gamma}\ddot{q}+\lambda_M(\hat{N}^\natural)'q - (\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})^{\cdot}\right)v\right]_{t=T} \\ &+ \int_0^T \left\{(\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})^{\cdot\cdot} - \left((\hat{\nu}+\lambda_M\hat{M}^\natural)\dot{q}+\hat{\gamma}\ddot{q}+\lambda_M(\hat{N}^\natural)'q\right)^{\cdot} + \lambda_M\hat{N}^\natural\dot{q}\right\}\cdot v\end{aligned} \quad (21)$$

As it often happens in variational calculus we proceed as follows:

1. Consider only the variations such that $v(T) = \dot{v}(T) = 0$. In this case $\psi'(0) = 0$ yields the following differential equations

$$\begin{aligned}&\hat{\mu}q^{(4)} + 2\dot{\hat{\mu}}q^{(3)} + (\ddot{\hat{\mu}} + \dot{\hat{\gamma}} - \hat{\nu} - \lambda_M\hat{M}^\natural)\ddot{q} \\ &\qquad + (\ddot{\hat{\gamma}} - \dot{\hat{\nu}} - \lambda_M(\dot{\hat{M}}^\natural + (\hat{N}^\natural)' - \hat{N}^\natural))\dot{q} - \lambda_M(\dot{\hat{N}}^\natural)'q = 0.\end{aligned} \quad (22)$$

2. Because of Eq. (21), $\psi'(0) = 0$ reduces to $\left[(\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})\dot{v} + \left(\hat{\nu}\dot{q}+\hat{\gamma}\ddot{q} - (\hat{\mu}\ddot{q}+\hat{\gamma}\dot{q})^{\cdot}\right)v\right]_{t=T} = 0$. Moreover, since $v(T)$ and $\dot{v}(T)$ can be chosen independent one of each other, then the vanishing of the first variation also implies that

$$\begin{aligned}&\hat{\mu}\ddot{q}(T) + \hat{\gamma}\dot{q}(T) = 0; \\ &-\hat{\mu}q^{(3)}(T) - \dot{\hat{\mu}}\ddot{q}(T) + (\hat{\nu} - \dot{\hat{\gamma}} + \lambda_M\hat{M}^\natural)\dot{q}(T) + \lambda_M(\hat{N}^\sharp)'q(T) = 0.\end{aligned} \quad (23)$$

We summarize the previous analysis in the statement of the following theorem:

Theorem F. *The Euler-Lagrange equation relative to the functional $\Gamma(q)$ defined on $\mathbf{X}$ are*

$$\hat{\mu}(t)q^{(4)}(t)+2\dot{\hat{\mu}}(t)q^{(3)}(t)+Z_2(t)\ddot{q}(t)+Z_1(t)\dot{q}(t)+Z_0(t)q(t)+\nabla_q\hat{U}(q,C) = 0. \quad (24)$$

where

$$Z_2 = \ddot{\hat{\mu}} + \dot{\hat{\gamma}} - \hat{\nu} - \lambda_M\hat{M}^\natural, \qquad Z_1 = \ddot{\hat{\gamma}} - \dot{\hat{\nu}} - \lambda_M(\dot{\hat{M}}^\natural + (\hat{N}^\natural)' - \hat{N}^\natural),$$

$$Z_0 = \hat{k} + \lambda_M\hat{O}^\natural - \lambda_M(\dot{\hat{N}}^\natural)', \quad (25)$$

together with the boundary conditions in Eq. (23).

It is worth mentioning that the above theorem holds also if we redefine $\Gamma(q)$ by arbitrary positive and bounded functions $\hat{\mu}(t)$, $\hat{\nu}(t)$, $\hat{\gamma}(t)$, and $\hat{k}(t)$. This is one of the key observations that made us devise a mechanism to deal with Eq. (23) without using the limiting procedure described in Chapter 2. In the following we will describe our initial ideas that we used to handle boundary conditions (23).

Boundary conditions. The solution of the forth-order differential equation on the filter parameters requires the satisfaction of the boundary conditions (23). The underlying idea that drives the learning process is that one is expected to solve the problem of determining the filters in a causal way, which corresponds with imposing Cauchy's initial condition. However, the solution of Eq. (24) under Cauchy's initial condition will not, in general, satisfy conditions (23) at the end of learning. Hence, we get into a dilemma that involves the choice of the initial conditions, since the values $q(T), \dot{q}(T), \ddot{q}(T), q^{(3)}(T)$ do depend on the video signal in $(0, T]$, that is on the "future." We can break the dilemma when pairing a couple of important remarks: First, a special case in which conditions (23) are satisfied is whenever we have still images at T, so as $N^\sharp = 0$, and

$$\dot{q}(T) = \ddot{q}(T) = q^{(3)}(T) = 0. \tag{26}$$

Second, without limitations of generality, the color field $C(x, t)$ in Π will always contain brief portions of null signal. Moreover, its eventual manipulation with the purpose of injecting brief portions of null signal does not change its information structure, so as one can reasonably regard the visual environment with such a manipulation equivalent with respect to the one from which it is generated. The intuition is that such a "reset" of the video results in $N^\sharp = 0$ and, moreover, the null signal also affects the differential equation of learning (24) by resetting the dynamics, so as $\dot{q}(T) = \ddot{q}(T) = q^{(3)}(T) = 0$ is also very well approximated. Hence, no matter what the initial conditions are, it turns out the we can satisfy conditions (23) after a small portions of the video.

Now, we will translate this intuition into a formal statements. Let us consider a sequence of times $0 < t_0 < t_1 < t_2 < \cdots < t_{2N} < T$ that defines the two sets $A = \bigcup_{i=0}^{N} A_i$ with $A_i = (t_{2i-1}, t_{2i})$, $t_{-1} = 0$ and $B = \bigcup_{i=0}^{N} B_i$ with $B_i = (t_{2i}, t_{2i+1})$, $t_{2N+1} = T$. Suppose furthermore that we modify the video signal in the following way $C(x, t) \to C(x, t)[t \in A]$, so that it is identically null on B. As already pointed out, in doing so, we do not change the problem of discovering visual features, since we just dilute the information that is contained in C. On the other hand, whenever $C = 0$, this results into a remarkable simplification of the system dynamics in B: the potential U and all the terms coming from the motion invariance term (the ones proportional to λ_M) are identically zero. Moreover, since the EL equations still holds true for time-variant coefficients $\hat{\mu}(t)$, $\hat{\nu}(t)$, $\hat{\gamma}(t)$, and $\hat{k}(t)$, we can always decouple the dynamics so that whenever $t \in B$ Eq. (24) becomes (see [7])

$$\bar{\mu} q^{(4)} + 2\bar{\theta}\bar{\mu} q^{(3)} + (\bar{\theta}^2 \bar{\mu} + \bar{\theta}\bar{\gamma} - \bar{\nu})\ddot{q} + (\bar{\theta}^2 \bar{\gamma} - \bar{\theta}\bar{\nu})\dot{q} + \bar{k} q = 0, \qquad t \in B \tag{27}$$

where $\bar{\theta}$, $\bar{\mu}$, $\bar{\nu}$ and $\bar{\lambda}$ are arbitrary constants different from θ, μ, ν and λ. In particular the following theorem guarantees us that $\bar{\theta}$, $\bar{\mu}$, $\bar{\nu}$ and $\bar{\lambda}$ can be chosen in such a way that the boundary conditions in Eq. (23) are approximaterly satisfied at the end of each B interval.

Theorem G. *We can always choose the system parameters of Eq.* (27) *in such a way that* $|q^{(k)}(t_{2i+1})| = 0$, $k = 1, 2, 3$, *up to an arbitrary precision for*

$i = 0, 1, \ldots, N$ regardless of the initial Cauchy conditions, which is in fact a special way of satisfying boundary conditions (23).

Proof. See [7] for the proof. ∎

The intuition behind this result is that the dynamical system defined by (27) becomes asymptotically stable under an appropriate choice of the parameters, which corresponds with driving the dynamics to a reset state arbitrarily fast.

Another important property of the dynamics in the B_i is that it we can arrange things in such a way that it does not alter the solution found in the previous A_j. More precisely, let $(0, \lambda_2, \lambda_3, \lambda_4)$ be the roots of the characteristic polynomial associated with Eq. (27) and let $V_3 = V(\lambda_2, \lambda_3, \lambda_4)$ be the Vandermonde matrix associated with the λ_i eigenvalues. The the following theorem holds.

Theorem H. *Let $\Lambda = (V(\lambda_2/\rho, \lambda_3/\rho, \lambda_4/\rho))^{-1}$ be and for every even $i = 0, \ldots, 2N$ consider the defined sets $A_i = (t_{i-1}, t_i)$, $B_i = (t_i, t_{i+1})$. It is always possible to choose the coefficients in Eq.* (27) *such that $\forall \epsilon > 0$, if we choose*

$$\rho > [(9C/\epsilon) \cdot \max_k |q^{(k)}(t_i)|]^{1/2} > 1 \tag{28}$$

we have $|q(t_{i+1}) - q(t_i)| < \epsilon$, where $|\Lambda_{kj}| \leq C$ for all k and $j = 1, 2, 3$.

Proof. See [7] for the proof. ∎

System dynamics. Here we will mainly focus on the "free dynamics" $C \equiv 0$. This turns out to be a good way to classify the solution of Eq. (24) according to the behaviour of the associated "free" equation (obtained for $C \equiv 0$) This case is particularly important since it is possible to analyze this case in details, and it gives us insights on the solutions depending on the choice of the parameters. Let $\chi(x) = x^4 + bx^3 + cx^2 + dx + e$ be the characteristic polynomial of the EL equation (24) with $U \equiv 0$ (which is just the same as Eq. (27) only with the unbarred variables); here we assume $\mu \neq 0$ and use the notation $b = 2\theta$, $c = (\theta^2\mu + \theta\gamma - \nu)/\mu$, $d = (\theta^2\gamma - \theta\nu)/\mu$, and $e = k/\mu$.

If we replace $x = z - b/4$ with $\chi(x)$ then we obtain the reduced quartic equation $\zeta(z) := \chi(z - b/4) = z^4 + qz^2 + rz + s = 0$, where $q = c - 3b^2/8, \quad r = b^3/8 - bc/2 + d, \quad s = b^2c/16 - 3/256b^4 - bd/4 + e$. Then one can prove (see [7]) that the following proposition holds:

Proposition D. *If we choose $\theta, \mu, \nu, \gamma_1, \gamma_2, k$ such that $\theta > 0$ and:*

$$\begin{aligned} &\mu > \gamma_2^2, \quad \nu > \gamma_1^2, \quad \nu < \theta\gamma_1\gamma_2, \quad 0 < k \leq \frac{(\nu - \theta\gamma_1\gamma_2)^2}{4\mu} \\ &\gamma_1 < 0, \quad \gamma_2 < \frac{\gamma_1}{\theta} \quad \text{or} \quad \gamma_1 > 0, \quad \gamma_2 > \frac{\gamma_1}{\theta}. \end{aligned} \tag{29}$$

then the following conditions are jointly verified:

1. Γ admits a minimum in $\mathbf{X}$;
2. the homogeneous equation associated with Eq. (27) has the following two properties:

i. it is asymptotically stable;
ii. it yields aperiodic dynamics (the roots of the characteristic polynomial are real).

Now we will present some experimental results of the theory.

3.2.1. Experimental Results

We implemented a solver for the differential equation of Eq. 3.2–(24) that is based on the Euler method. After having reduced the equation to the first order, the variables that are updated at each time instant are q, $\dot{q}$, $\ddot{q}$, and $q^{(3)}$. The code and data we exploited to run the following experiments can be downloaded at `http://www.dii.unisi.it/ melacci/calneco/neco_code_data_params.zip` together with the full list of model parameters.

We randomly selected two real-world video sequences from the Hollywood Dataset HOHA2 `marszalek09`, that we will refer to as "skater" and "car", and a clip from the movie "The Matrix" (©Warner Bros. Pictures). The frame rate of all the videos is ≈ 25 fps, so we set the step-size of the Euler method to $1/25$, and each frame was rescaled to 240×110 and, for simplicity, it was converted to grayscale. Videos have different lengths, ranging from ≈ 10 to ≈ 40 seconds, and they were looped until $45,000$ frames were generated, thus covering a significantly longer time span.

We randomly initialized the variable q for $t = 0$, while the derivatives $\dot{q}$, $\ddot{q}$, and $q^{(3)}$ were set to 0. We used the softmax function to force a probabilistic activation of the features, as suggested in Section Eq. 3.2–(33), and we computed the optical flow v using an implementation from the OpenCV library. Convolutional filters cover squared areas of the input frame, and we set g_x, for each x, to be the inverse of the frame area, i.e., we assume that we have a uniform distribution over the retina. All the results that we report are averaged over 10 different runs of the algorithms.

The video is presented gradually to the agent so as to favour the acquisition of small chunks of information. We start from a completely null signal (all pixel intensities are zero), and we slowly increase the level of detail and the pixel intensities, in function of $\tau(t) \in [0, 1]$, where $\tau(t) = 0$ leads to null signal and $\tau(t) = 1$ to full details. In detail,

$$C(x, t) = \tau(t) G_{(1-\tau(t))\delta} * \overline{C}(x, t), \tag{1}$$

where $*$ is the spatial convolution operator, $\overline{C}(x, t)$ is the original source video signal, G_σ is a Gaussian filter of variance σ, and $\delta > 0$ is a customizable scaling factor, that we set to the size of the squared discrete Gaussian filter mask. It is easy to see that for $\tau(t) = 1$ we get $C(x, t) = \overline{C}(x, t)$. We start with $\tau(0) = 0$, and then $\tau(t),\ t > 0$, is progressively increased as time passes with the following rule,

$$\tau(t + 1) = \tau(t) + \eta(1 - \tau(t)), \tag{2}$$

where we set $\eta = 0.0005$. We refer to the quantity $1 - \tau$ as the "blurring factor", being it proportional to the variance of the Gaussian blur.

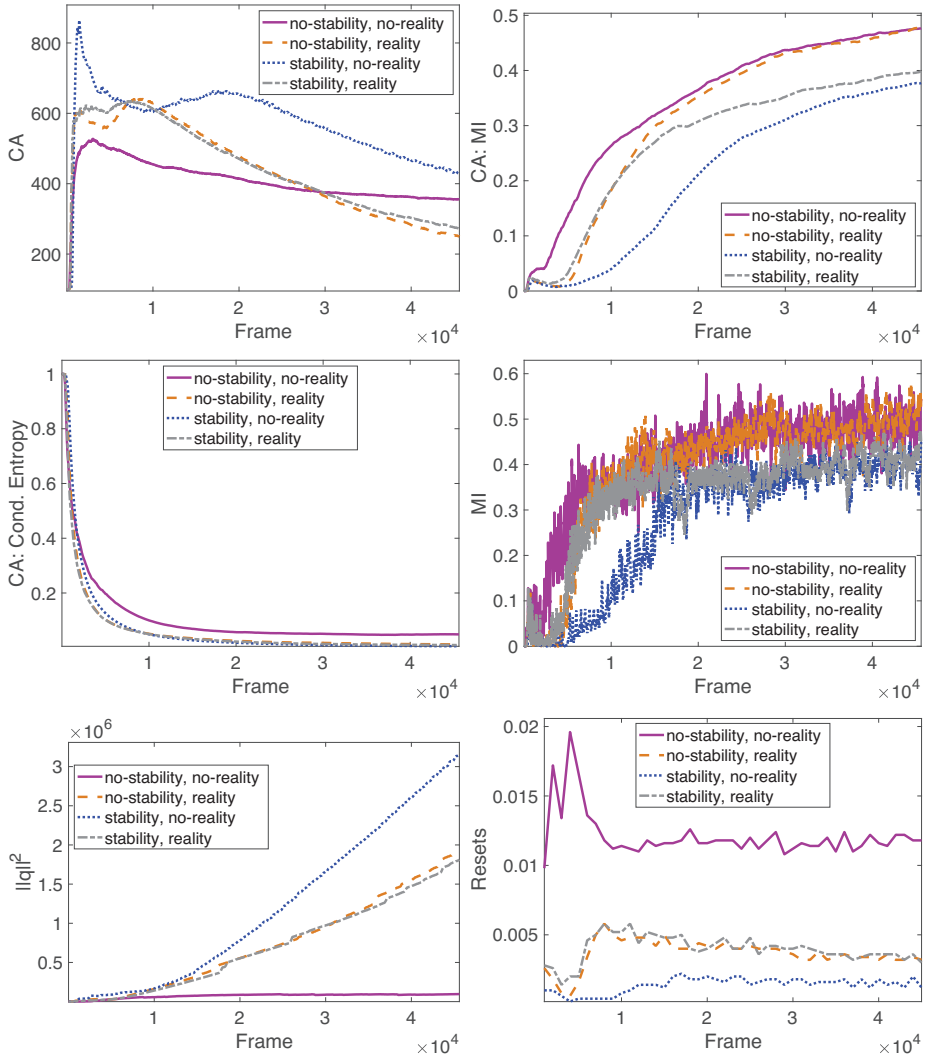


Fig. 12. Comparing 4 configurations of the parameters, characterized by different properties in terms of stability and reality of the roots of the characteristic polynomial. The input video is reproduced (in loop) for 45k frames (x-axis). From left-to-right, top-to-bottom we report the Cognitive Action (CA), the portion of the cognitive action that is about the Mutual Information (MI) (that we maximize), the portion that is about the Conditional Entropy, the MI per-frame, the norm of $q(t)$, and the fraction of "reset" operations performed every 1000 frames.

In order to be able to (approximately) satisfy the conditions in Eq. 3.2–(23) we need to keep the derivatives small, so we implement a "reset plan" according to which the video signal undergoes a reset whenever the derivatives become too large. Formally, if $|\dot{q}(t')|^2 \geq \epsilon_1$, or $|\ddot{q}(t')|^2 \geq \epsilon_2$, or $|q^{(3)}(t')|^2 \geq \epsilon_3$ then we set to

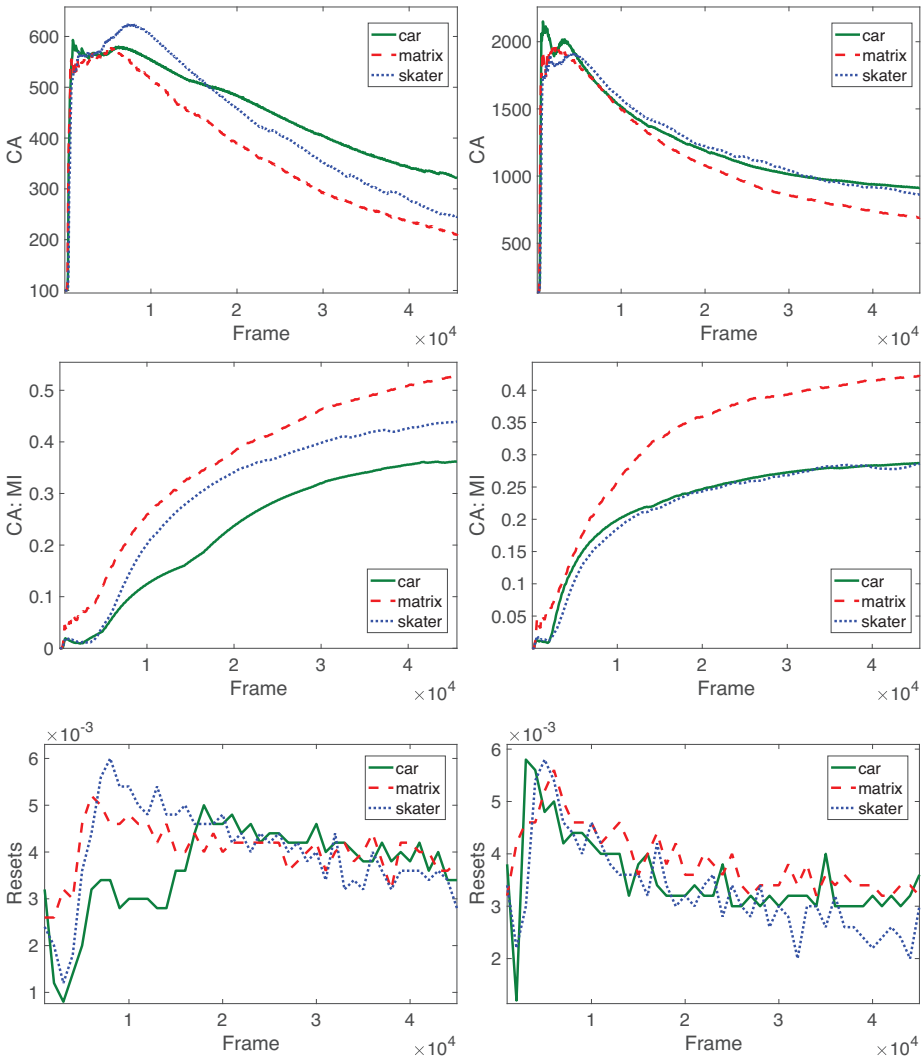


Fig. 13. Different number of features and filter sizes (1st column: $n = 5, size = 5 \times 5$; 2nd column: $n = 11, size = 11 \times 11$) in 3 videos. See Fig. 12 for a description of the plots.

0 all the derivatives, and we also force $\tau(t')$ to 0, leading to null video signal, as described above. We used $\epsilon_j = 300n$, for all j.

Our experiments are designed (i) to evaluate the dynamics of the cognitive action in function of different temporal regularities imposed to the model weights (parsimony), and then (ii) to evaluate the effects of motion, that introduces a spatio-temporal regularization on single and multi-layer architectures. When evaluating the temporal regularities, the cognitive action is composed by the

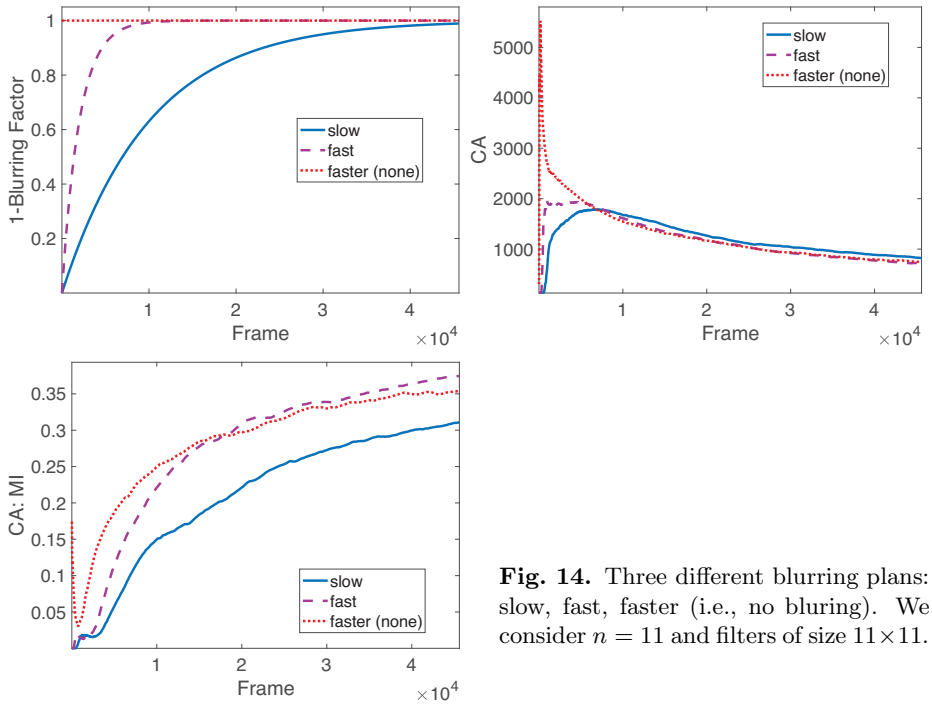


Fig. 14. Three different blurring plans: slow, fast, faster (i.e., no bluring). We consider $n = 11$ and filters of size 11×11.

information-based and parsimony terms only, and we experiment four instances of the set of parameters $\{\mu, \nu, \gamma, k\}$ of Eq. 3.2–(14), leading to different dynamics. Each instance is characterized by the roots of the characteristic polynomial that lead to *stable* or *not-stable* configurations, and with only *real* or also *imaginary* parts, keeping the roots close to zero, and fulfilling the conditions of Proposition D when stability and reality are needed. These configurations are all based on values of $k \in [10^{-19}, 10^{-3}]$, while $\theta = 10^{-4}$.

We performed experiments on the "skater" video clip, setting $n = 5$ features, and filters of size 5×5. Results are reported in Fig. 12. The plots indicate that there is an initial oscillation that is due to the effects of the blurring factor, that vanish after about 10k frames. The Mutual Information (MI) ($\mathcal{I}$) portion of the cognitive action correctly increases over time, and it is pushed toward larger values in the two extreme cases of "no-stability, reality" and "no-stability, no-reality". The latter shows more evident oscillations in the frame-by-frame MI value, due to the roots with imaginary part. In all the configurations the norm of q increases over time (with different speeds), due to the small values of k, while the frequency of reset operations is larger in the "no-stability, no-reality" case, as expected.

We evaluated the quality of the developed features by freezing the final q of Fig. 12 and computing the MI index over a single repetition of the whole video clip, reporting the results in Tab. 1–(a). This is the procedure we will follow in

Table 1
MUTUAL INFORMATION

(a)		(b)			(c)	
Config	(Skater)	Video	$n = 5, 5 \times 5$	$n = 11, 11 \times 11$	Blurring	$(n = 10, 5 \times 5)$
$\overline{S}\,\overline{R}$	0.54 ± 0.07	Car	0.38 ± 0.03	0.272 ± 0.003	Slow	0.35 ± 0.08
$\overline{S}\,R$	0.54 ± 0.08	Matrix	0.60 ± 0.03	0.45 ± 0.02	Fast	0.39 ± 0.05
$S\,\overline{R}$	0.44 ± 0.11	Skater	0.45 ± 0.13	0.35 ± 0.05	None	0.34 ± 0.08
$S\,R$	0.45 ± 0.13					

MI on (a) the "skater" video, given the models of Fig. 12 (S=stability, R=reality, $\bar{X}$=not X); (b) different videos, number of features, filter sizes (SR); (c) different blurring plans (SR)

the rest of the section when reporting numerical results in all the tables. We notice that, while in Fig. 12 we compute the MI on a frame-by-frame basis, here we compute it over the whole frames of the video at once, thus in a batch-mode setting. The result confirms that the two extreme configurations "no-stability, reality" and "no-stability, no-reality" show better results, on average. These performances are obtained thanks to the effect of the reset mechanism, that allows even such unstable configurations to develop good solutions. When the reset operations are disabled, we easily incurred into numerical errors due to strong oscillations while, for example, the "stability" cases were less affected by this phenomenon.

We also compared the dynamics of the system on multiple video clips and using different filter sizes (5×5 and 11×11) and number of features ($n = 5$ and $n = 11$) in Fig 13. We selected the "stability, reality" configuration of Fig. 12, that fulfils the conditions of Proposition D. Changing the video clip does not change the considerations we did so far, while increasing the filter size and number of features can lead to smaller MI index values, mostly due to the need of a better balancing the two entropy terms to cope with the larger number of features. The MI of Tab. 1–(b) confirms this point. Interestingly, the best results are obtained in the longer video clip ("The Matrix") that requires less repetitions of the video, being closer to the real online setting.

Figure 14 and Tab. 1–(c) show the results we obtain when using different blurring plans ("skater" clip), that is, different values of η in Eq. (2), that lead to the blurring factors reported in the first graph of Fig. 14. These results suggest that a gradual introduction of the video signal helps the system to find better solutions than in the case in which no-plans are used, but also that a too-slow plan is not beneficial. The cognitive action has a big bump when no-plans are used, while this effect is more controlled and reduced in the case of both the slow and fast plans.

In order to study the effect of motion in multi-layer architectures (up to 3 layers), we still kept the most stable configuration ("stability, reality", 5×5 filters, 5 features), and introduced the motion-related term in the cognitive action. Our multi-layer architecture is composed of a stack of computational models developed accordingly to (3). A new layer ℓ is activated whenever layer $\ell - 1$ has processed a large number of frames ($\approx 45k$), and the parameters of

Table 2
MUTUAL INFORMATION–LAYERS

		$\lambda_M = 0$	10^{-8}	10^{-6}	10^{-4}	10^{-2}	1	10^2
Skater	$\ell = 1$	$.61 \pm .11$	$.54 \pm .11$	$.52 \pm .07$	$.53 \pm .08$	$\mathbf{.69} \pm .07$	$.53 \pm 0$	$.01 \pm 0$
	$\ell = 2$	$.53 \pm .12$	$\mathbf{.62} \pm .15$	$.60 \pm .11$	$.43 \pm .06$	$.48 \pm .06$	$.1 \pm .1$	$.03 \pm .01$
	$\ell = 3$	$.56 \pm .17$	$.58 \pm .20$	$\mathbf{.62} \pm .10$	$.18 \pm .16$	$.16 \pm .17$	$.04 \pm .02$	$.03 \pm .02$
Car	$\ell = 1$	$.49 \pm .05$	$.44 \pm .02$	$.46 \pm .04$	$.47 \pm .04$	$\mathbf{.66} \pm .10$	$.60 \pm .02$	$.01 \pm 0$
	$\ell = 2$	$.25 \pm .26$	$.54 \pm .10$	$\mathbf{.65} \pm .08$	$.46 \pm .03$	$.63 \pm .11$	$.18 \pm .32$	$.03 \pm .01$
	$\ell = 3$	$.26 \pm .34$	$.45 \pm .22$	$\mathbf{.51} \pm .11$	$.38 \pm .20$	$.24 \pm .20$	$.09 \pm .12$	$.04 \pm .02$
Matrix	$\ell = 1$	$.66 \pm .01$	$.66 \pm .02$	$\mathbf{.67} \pm .01$	$.63 \pm .05$	$.59 \pm .03$	$.44 \pm 0$	$.23 \pm .02$
	$\ell = 2$	$.55 \pm .13$	$.56 \pm .14$	$.43 \pm 0$	$.45 \pm .04$	$\mathbf{.62} \pm .02$	$.35 \pm .19$	$.13 \pm .08$
	$\ell = 3$	$\mathbf{.64} \pm .03$	$.54 \pm .11$	$.35 \pm .07$	$.40 \pm .01$	$.21 \pm .07$	$.06 \pm .03$	$.04 \pm .02$

MI in different videos, up to 3 layers ($\ell = 1, 2, 3$), and for multiple weighting factors λ_M of the motion-based term. All layers share the same λ_M.

Table 3
MUTUAL INFORMATION–LAYERS 2

		$\lambda_M = 0$	10^{-8}	10^{-6}	10^{-4}	10^{-2}	1	10^2
Skater	$\ell = 1$	$.61 \pm .11$	$.54 \pm .11$	$.52 \pm .07$	$.53 \pm .08$	$\mathbf{.69} \pm .07$	$.53 \pm 0$	$.01 \pm 0$
	$\ell = 2$	$.38 \pm .34$	$\mathbf{.53} \pm .12$	$.50 \pm .1$	$.47 \pm .1$	$.41 \pm .02$	$.33 \pm .17$	$.21 \pm .2$
	$\ell = 3$	$.55 \pm .12$	$\mathbf{.62} \pm .11$	$.55 \pm .13$	$.42 \pm .01$	$.36 \pm .09$	$.2 \pm .18$	$.39 \pm .22$
Car	$\ell = 1$	$.49 \pm .05$	$.44 \pm .02$	$.46 \pm .04$	$.47 \pm .04$	$\mathbf{.66} \pm .10$	$.60 \pm .02$	$.01 \pm 0$
	$\ell = 2$	$.48 \pm .1$	$.59 \pm .17$	$.59 \pm .18$	$.55 \pm .12$	$.41 \pm .01$	$.01 \pm 0$	$\mathbf{.64} \pm .01$
	$\ell = 3$	$.67 \pm .01$	$.60 \pm .12$	$\mathbf{.73} \pm .09$	$.36 \pm .05$	$.33 \pm .11$	$.27 \pm .14$	$\mathbf{.73} \pm .01$
Matrix	$\ell = 1$	$.66 \pm .01$	$.66 \pm .02$	$\mathbf{.67} \pm .01$	$.63 \pm .05$	$.59 \pm .03$	$.44 \pm 0$	$.23 \pm .02$
	$\ell = 2$	$.55 \pm .13$	$.56 \pm .14$	$.43 \pm 0$	$.45 \pm .04$	$\mathbf{.62} \pm .02$	$.35 \pm .19$	$.13 \pm .08$
	$\ell = 3$	$.55 \pm .12$	$.53 \pm .12$	$\mathbf{.82} \pm .14$	$.35 \pm .05$	$.35 \pm .31$	$.02 \pm .01$	$.01 \pm 0$

Same structure of Tab. 2. Here the model with the best λ_M is selected and used as basis to activate a new layer (layer $\ell = 1$ is the same as Tab. 2)

layer $\ell - 1$ are not updated anymore. We initially considered the case in which all the layers $\ell = 1, \ldots, 3$ share the same value λ_M that weighs the motion-based term. Tab. 2 shows the MI we get for different weighting schemes. Introducing motion helps in almost all the cases (for appropriate λ_M — the smallest values of λ_M are a good choice on average), and, as expected, a too strong enforcement of the motion-related term leads to degenerate solutions with small MI. We repeated these experiments also in a different setting. In detail, after having evaluated layer ℓ for all the values of λ_M, we selected the model with the largest MI and started evaluating layer $\ell + 1$ on top of it. Tab. 3 reports the outcome of this experience. We clearly see that motion plays an important role in increasing the average MI. In the case of "car", we also obtained two (unexpected) positive results when strongly weighing λ_M. They are due to very frequent reset operations, that avoid the system to alter the filters when the too-strongly-enforced motion-based term yields very large derivatives. This is an interesting behaviour that, however, was not common in the other cases we reported.

3.3. FURTHER DEVELOPMENTS OF THE THEORY

In Section 3.1 we discussed the principles on which our theory of vision stands, while in Section 3.2 we proposed a specific model that we derived from the general theory under a number of assumptions, some of which were strongly motivated and suggested by mathematical results, some by computational issues while others were assumed as working hypotheses. In this section we want, on the basis of what we learned either from the theoretical analysis presented in Section 2.2 and also from the experimental results shown in Section 3.2.1, to reconsider some of these working hypotheses. In particular we will re-discuss two main points of the theory that we believe that are crucial and that could result in major improvement on the model developed in Section 3.2:

1. The motion invariance term based on the trajectory of the focus of attention (as discussed in item ii. at the end of Section 3.1);
2. Instead of the "reset mechanism" that we introduces in the last section to handle the boundary conditions, applying the De Giorgi approach to the problem as it is discussed on Chapter 2.

A closing section will then wrap the discussion up on two important issues that will be the core part of future investigations: Deep architectures and integrated supervisions.

Let us start to consider the implications of 1. in terms of the Lagrangian theory presented in Section 3.2.

3.3.1. Motion Term Driven by the Focus of Attention

In Section 3.1.1 we defined the motion invariance term according to Eq. 3.1.1–(13). This constraint was then incorporated in the theory by adding to the Lagrangian a penalty term integrated over the whole retina. However we soon realized that this way of imposing the motion invariance term would only result in a mean effect on the features: Since the features that we are extracting are convolutional the filters are shared weights all over the retina, thus if we try to force coherence under motion on all the pixels what we will end up with is just a mean, regularizing effect. On the other hand the intuition is that if we focus on a single pixel, then we can hope to be able to change the convolutional filters in such a way to have motion-coherent feature at least where the attention is focused.

Another great advantage of using motion invariance only on the focus of attention, although in a certain sense can be considered as a special case of the above comment, is the fact that brakes down any sort of ambiguities that could arise for example for the presence of multiple version of the same shapes at different scales or moving on different trajectories.

Sot let us see how the motion invariance term on the convolutional activations looks like when computed on the trajectory of the focus of attention $a(t)$:

$$\frac{d}{dt}A_i(a(t),t) = \partial_t A_i(a(t),t) + \nabla A_i(a(t),t)\cdot \dot{a}(t). \tag{1}$$

The discrete counterpart of this is*

$$\frac{d}{dt}A_{ia(t)} = \frac{d}{dt}(\chi^i_\alpha \gamma_\alpha(a(t),t)) = \dot{\chi}^i_\alpha(t)\gamma_\alpha(a(t),t) + \chi^i_\alpha(t)\frac{d}{dt}\gamma_\alpha(a(t),t) = 0. \quad (2)$$

Hence the motion invariance terms becomes

$$\begin{aligned}\int_0^T \varpi(t)\frac{dA_{ia(t)}}{dt}\frac{dA_{ia(t)}}{dt}\,dt = \int_0^T \varpi(t)\Big(&\dot{\chi}^i_\alpha(t)\gamma_\alpha(a(t),t)\gamma_\beta(a(t),t)\dot{\chi}^i_\beta(t)\\ &+ 2\chi^i_\alpha(t)\frac{d}{dt}\gamma_\alpha(a(t),t)\gamma_\beta(a(t),t)\dot{\chi}^i_\beta(t)\\ &+ \chi^i_\alpha(t)\frac{d}{dt}\gamma_\alpha(a(t),t)\frac{d}{dt}\gamma_\beta(a(t),t)\chi^i_\beta(t)\Big)\,dt.\end{aligned} \quad (3)$$

As we did in Section 3.2 we can define the matrices (that for convenience we will indicate with the same names as before) $O_{\alpha\beta} := (d\gamma_\alpha(a(t),t)/dt)\,(d\gamma_\beta(a(t),t)dt)$, $N_{\alpha\beta} := (d\gamma_\alpha(a(t),t)/dt)\gamma_\beta(a(t),t)$ and $M_{\alpha\beta} := \gamma_\alpha(a(t),t)\gamma_\beta(a(t),t)$. Using the same exact definition that we have given in Section 3.2 of $M^\natural$, $N^\natural$, $O^\natural$ and of q we finally get the expression for the motion term:

$$\int_0^T \varpi(t)\left(\frac{1}{2}\dot{q}M^\natural(t)\dot{q} + qN^\natural(t)\dot{q} + \frac{1}{2}q(t)O^\natural(t)q(t)\right)\,dt. \quad (4)$$

which is formally the same as the one in Eq. 3.2–(11). In this expressions however the matrices $M^\natural$, $N^\natural$, $O^\natural$ depends on time other than through the video also directly through the trajectory of the focus of attention and its derivative.

It is also interesting to notice that the same results could have been found by choosing $g_x \equiv \delta_{xa(t)}$ so that all the sum on the retina that appeared in the old definition of the matrices $M^\natural$, $N^\natural$, $O^\natural$ would have collapsed on the focus of attention.

3.3.2. Causal Formulation

We will now apply the techniques described in Chapter 2 to our theory of vision. In particular we will try to reformulate the theory presented in Section 3.2 using the De Giorgi causal approach. In Section 2.2.2 we investigated the possibility of using a sequential optimization for problems which have explicit temporal dependence in the potential. We also argued that this approach is particularly fruitful when we have a natural way to segment the entire agent life in temporal segments inside of which we have consistent information. In a visual system with an attention mechanism $a(t)$ which behaves as a natural (human) focus of attention a possible segmentation criterion as intended in Section 2.2.2 is furnished by the attention signal itself as we can choose the times of the temporal partition t_i to coincide with the instants in which the attention quickly moves from a region of the retina to another (saccadic movements).

* Here we will use a slightly different notation than in in Section 3.2 for the quantity γ; what we called $\gamma^x(t)$ here will be denoted as $\gamma(x,t)$, both quantities however have the same definition.

This being said consider the following family of functionals for the convoluitional filters q:

$$F_\varepsilon(q) := \int_0^T e^{-t/\varepsilon}\Big(\varepsilon^2\frac{\rho}{2}|\ddot{q}|^2 + \varepsilon\frac{\nu}{2}|\dot{q}|^2 + \lambda_M\Big(\frac{f_M(\varepsilon)}{2}\dot{q}\cdot M^\natural\dot{q} + f_N(\varepsilon)q\cdot N^\natural\dot{q} + \frac{f_O(\varepsilon)}{2}q\cdot O^\natural q\Big) + U(q,t)\Big)\,dt, \tag{1}$$

where the terms f_M, f_N and f_O are suitable positive weights such as the ε^2 and ε in front of the regularization terms that we reserve to choose appropriately on the basis of the resulting Euler equations.

Using exactly the same arguments as in Section 2.1.2 (Theorem 2.1.2A), we can conclude that if the U, $M^\natural$, $N^\natural$ and $O^\natural$ are smooth functions, then we gain the regularity of the minimum of the functional F_ε (whose existence we have basically proved when we proved the existence of the minimum for Γ in Section 3.2). Therefore we can write down the Euler equations for this functional; we have

$$\begin{aligned} &e^{-t/\varepsilon}\rho\varepsilon^2 q^{(4)} - e^{-t/\varepsilon}\frac{2}{\varepsilon}\varepsilon^2\rho q^{(3)} + e^{-t/\varepsilon}\Big(\frac{1}{\varepsilon^2}\varepsilon^2\rho - \varepsilon\nu - \lambda_M f_M(\varepsilon)M^\natural\Big)\ddot{q}\\ &+e^{-t/\varepsilon}\Big(\frac{1}{\varepsilon}\varepsilon\nu - \lambda_M\big(f_M(\varepsilon)(\dot{M}^\natural - \frac{1}{\varepsilon}M^\natural) + f_N(\varepsilon)(N^{\natural\prime} - N^\natural)\big)\Big)\dot{q}\\ &+e^{-t/\varepsilon}\frac{\lambda_M f_N(\varepsilon)}{\varepsilon}N^{\natural\prime}q - e^{-t/\varepsilon}\lambda_M f_N(\varepsilon)\dot{N}^{\natural\prime}q + e^{-t/\varepsilon}\lambda_M f_O(\varepsilon)O^\natural q + \nabla U = 0. \end{aligned} \tag{2}$$

As usual the Euler equations (see Appendix A) are coupled with two conditions at $t = T$ which are $L_a(T, q(T), \dot{q}(T), \ddot{q}(T)) = 0$ and $L_p(T, q(T), \dot{q}(T), \ddot{q}(T)) - dL_a(T, q(T), \dot{q}(T), \ddot{q}(T))/dt = 0$:

$$\begin{aligned} &e^{-t/\varepsilon}\varepsilon^2\rho\ddot{q}(T) = 0;\\ &-e^{-t/\varepsilon}\varepsilon^2\rho q^{(3)}(T) + e^{-t/\varepsilon}\frac{1}{\varepsilon}\varepsilon^2\rho\ddot{q}(T) + e^{-t/\varepsilon}(\varepsilon\nu + \lambda_M f_M(\varepsilon)M^\natural)\dot{q}(T)\\ &\qquad + e^{-t/\varepsilon}\lambda_M f_N(\varepsilon)N^{\natural\prime}q(T) = 0. \end{aligned} \tag{3}$$

Looking at Eq. (2) we immediately realize that $f_M(\varepsilon) = \varepsilon$ is a good choice, indeed if we were to choose $f_M(\varepsilon) = \varepsilon^n$ with $n < 1$ formally the Euler equation, as $\varepsilon \to 0$ would reduce to something, which would depend on the mutual value of f_M, f_N and f_O but not on ∇U and neither on any other terms that comes from the regularization part. The same remark also hold for $f_N(\varepsilon)$. So far then it seems that a good choice would be $f_M(\varepsilon) = f_N(\varepsilon) = \varepsilon$. Now for the choice of $f_O(\varepsilon)$ we are left with two possibilities, indeed if we were to choose $f_O(\varepsilon) = \varepsilon^n$ with $n < 0$ we would get (after having chosen the other two terms to be order ε) an even more degenerate equation. Then we can either choose $n > 0$ or $n = 0$; in the first case the Euler equations will formally converge as $\varepsilon \to 0$ to

$$\rho\ddot{q} + (\nu + \lambda_M M^\natural)\dot{q} + \lambda_M N^{\natural\prime}q + \nabla U = 0. \tag{4}$$

Instead with the choice $n = 0$ we would get

$$\rho\ddot{q} + (\nu + \lambda_M M^\natural)\dot{q} + \lambda_M N^{\natural\prime} q + \lambda_M O^\natural q + \nabla U(q,t) = 0. \tag{5}$$

Both these limiting equations seems to be reasonable even though one would say that Eq. (5) is much more in the spirit that we described in Chapter 2 since we can always regard the term $O^\natural$ a s a part of the potential.

With this choices the functional F_ε becomes:

$$F_\varepsilon(q) := \int_0^T e^{-t/\varepsilon}\Big(\varepsilon^2\frac{\rho}{2}|\ddot{q}|^2 + \varepsilon\frac{\nu}{2}|\dot{q}|^2 + \lambda_M\Big(\frac{\varepsilon}{2}\dot{q}\cdot M^\natural\dot{q} + \varepsilon q\cdot N^\natural\dot{q} + \frac{1}{2}q\cdot O^\natural q\Big) + U(q,t)\Big)\,dt. \tag{6}$$

Using this functional we can lay down the following strong conjecture based on the formal limit that led to Eq. (5)

Conjecture C. *For $\rho > 0$ the solution of* (2) *with boundary conditions* (3) *and Cauchy initial conditions $q(0) = q^0$, $\dot{q}(0) = q^1$ converges to the solution of* (5) *with the same boundary conditions.*

Notice that the time dependent matrices $M^\natural$, $N^\natural$, $O^\natural$ can always be chosen to be L^∞. This observation is central as it suggests that we can replace the fourth order dynamics 3.2–(24) and our "reset plan" with a much simpler second order ODE solved with Cauchy initial conditions. And there is no doubt that a simpler structure of the differential equation also opens the door to the possibility to perform some analysis directly on the differential equations in order to better understand the nature of the leaning process.

For example when we force the motion invariance terms (i.e. we choose consider the regime of big λ_M) formally we have that the dynamics is mainly driven by the term $\lambda_M M^\natural\dot{q} + \lambda_M N^{\natural\prime} q + \lambda_M O^\natural q$. Now if we look back at the definition of the matrices we soon realize that when we are focusing on a coherent part of an image the term $\gamma_\alpha(a(t),t)$ is slowly changing since in the neighbour of the focus of attention we will get similar values of the video. Therefore the entries of the matrix $O^\natural$ will be negligible compared with the other two terms. In this approximation the requirement $\lambda_M M^\natural\dot{q} + \lambda_M N^{\natural\prime} q = 0$ is equivalent to the satisfaction of the motion invariance constraint 3.3.1–(1) (just multiply this relation by $\dot{q}$).

3.3.3. Multilayer Architecture and Supervisions: a Prospective

In the experimental Section 3.2.1 we discussed the possibility of defining a *progressive* multilayer architecture. What we did was to use the single layer theory to sequentially (starting from the lower layers to the higher ones) develop the convolutional filters. Of course this solution is a zeroth order approximation of a proper multilayer theory since it lacks of any kind of backpropagation from higher levels.

We can, of course, formulate the theory taking into account a multilayer architecture from the beginning by explicitly writing the inter-layers dependence

in the functionals that we use to define the theory. Indeed the terms of the potential will result in a function of all the convolutional filters with a rich compositional structure that can in principle be handled in the Euler Lagrange equations with a backprop-like calculation. However we will encounter more severe problems with the multilayer generalization of the motion invariance term.

Suppose that we want to impose a motion invariance term on the convolutional activations of each layer. These terms will in general depend on the convolutional filters of that same level and on its derivative, but it will also essentially depend on the variables at lower levels *and* on their derivatives though the value of the feature map on which the invariance is imposed. This, of course, makes the problem of computing the variations of the functional intractable; even we just two layers the calculations are extremely involved.

A possible solution to this apparent impasse is to exploit the power of the variational formulation that we are adopting and to apply the theory for architectural constraints that we have developed in Section 2.3 (actually our interest in that kind of formalism was originally inspired originally exactly by this problem).

If we do so the complete multilayer formulation will consist of a variational problem on the variables of all the different layers which will simply consists of the sum of the functionals for each layer together with a set of constraints that are exactly of the form that we discussed in Section 2.3.2.

Another aspect of the theory that we want to emphasize is the possibility to incorporate (both in the one presented in Section 3.2 and the one proposed at the beginning of this section) supervision in a uniform and simple way. In particular suppose that we want to incorporate a set of supervised examples that comes at times $t_0, t_1, t_3, \dots, t_K$; this can be done by adding to the potential that we have used so far in vision (that consists essentially of terms that come from the mutual information) a new potential that makes memory of the given examples. We have already discussed a possibility for such a potential when discussing the transposition of the theory of Poggio and Girosi [44] in the temporal domain with the potential defined in Eq. 1.2.2–(4). Another possibility would be the following: Suppose that $v_0(\omega)$ is the loss relative to the example presented at t_0 and in general $v_i(\omega)$ the loss relative to the example presented at time t_i (for example if at time $t = t_5$ the following pair example-supervision is given (x_5, y_5) we may define $v_5 = 1/2(y_5 - f_\omega(x_5))^2$ where f is the overall model that should map the examples into the predictions) then we define

$$U_S(\omega, t) := \begin{cases} v_0(\omega) & \text{for } t_0 \geq t < t_1; \\ v_0(\omega) + v_1(\omega) & \text{for } t_1 \geq t < t_2; \\ \quad \vdots & \\ \sum_{i=0}^{k} v_i(\omega) & \text{for } t \geq t_k. \end{cases}$$

This potential indeed accumulates the various examples and keeps memory of the supervised example that the agent saw during its life. Notice that for $t > t_K$ the potential lose its temporal dependence and becomes a constant risk.

CONCLUSIONS

My conclusions have cost me some labor from the want of coincidence between accounts of the same occurrences by different eyewitnesses, arising sometimes from imperfect memory, sometimes from undue partiality for one side or the other.

— THUCYDIDES (*Peloponnesian War, I, 22)*

In this work we tried to pursue the coherent idea of learning processes as phenomena which takes place in time. In particular we showed how batch learning problems can be reformulated using time-independent potential functions like it happens in many problems of classical mechanics. We showed that in this case the resulting equations of motion generalizes batch learning optimized with gradient descent. We argued how the case of time-dependent potentials, instead, is much more in the spirit of an online stochastic method.

We discussed how the explicit temporal dependencies of the potential are of the utmost importance whenever we want to regard learning processes as an interplay between the internal dynamics of the learner and the temporal scale at which the input signal (which carry the data relevant for the task in question) varies. It is in this contest that we firstly see the importance of dissipative dynamics, which is responsible of relaxation towards states that are as good as possible in terms of the learning task we are trying to solve. We also showed that the dynamical behaviour of the evolution laws borrowed from classical mechanics are well suited for learning: When paired with the the additional hypothesis that the temporal dependence of the potential consists of a signal which has some properties of periodicity (quasi-periodicity), we can observe how eventually also the prediction of the model will share the same pattern of repetition so to have consistent predictions on similar data.

We show how both formulations, time-independent (batch mode) and time-dependent, can be recovered though a variational formulation that allow us to get a much more high level control of the theory. This becomes particularly apparent in the theory of visual extraction. The motion invariance term is a perfect example of a class of constraints that act on the dynamics and that when introduced in the Lagrangian it cannot be classified neither as a potential term neither as a kinetic term. Therefore its contribution to the updates rules for the filters could not have been easily guessed otherwise.

A central part of the discussion was about the causality issues that arises when trying to formulate an evolution problem using calculus of variations. With this respect we used recent results that show that causality can be conquered at

the price of introducing both an overall weight to the Lagrangian and a temporal scale that in some sense measure the "degree of causality" of the solution. We saw how this causal solution may be more satisfactorily interpreted in terms of a number of sequential, independent variational problems performed on consistent data. We also showed how the inter-layer dependencies of a feedforward NN can be completely described by means of holonomic constraints; in particular we described how the lagrange multipliers for these constraints can be expressed in terms of the other variables of the problem.

The last part of the work is dedicated to the development of a theory of extraction of visual feature. The formulation of such a problem has been carried out using again methods from calculus of variations. Under the hypothesis of using convolutional features we showed how the computational soundness of the theory naturally requires an hypothesis of receptive fields. Moreover, as we have already pointed out this general approach has given us the possibility of including a dynamical constraint on the convolutional filters: The motion invariance constraint. In particular we argued how this term becomes very important when we can take advantage of a method for focusing attention.

To sum up the most important contributions of this dissertation are:

- Interpretation of batch mode learning and online learning in terms potential functions. We completed this discussion with a consistency result on classification of similar patterns (Theorem 1.2.3B);
- Formulation of learning as a variational problem, causality of the obtained solution and its interpretation in terms of a sequence of independent variational problems;
- Description of a FNN architecture in terms of holonomic constraints and consistency check with Backpropagation;
- Definition of a theory for convolutional feature extraction based on a functional index that enforces the dynamical constraint of motion invariance;
- Discussion of the associated Euler equations and how they suggest receptive-field-like solutions and the need for a focus of attention mechanism both on a continuous and on a discretized retina.
- Causality issues in vision and subsequent reformulation of the theory with the methods discussed in Chapter 2.

In the last section we lay down the guidelines for future and necessary developments of the theory: A proper multilayer architecture and an integrated way to inject supervision. The multilayer architecture is particularly important because of our assumptions on receptive fields while the need for a good way to include supervised example is crucial in view of more sophisticated interaction protocol between the learning agents and humans.

APPENDIX A

CALCULUS OF VARIATIONS

You will find it a very good practice always to verify your references, sir.

— MARTIN JOSEPH ROUTH, from J. W. Burgon, *Memoir of Dr. Routh* (1878)

CALCULUS OF VARIATION in its essence is the study of extremals of functions $f\colon \mathbf{X} \to \overline{\mathbf{R}}$, where $\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, +\infty\}$. The case in which $\mathbf{X}$ is a Euclidean space corresponds of course to the study of stationary points of a real valued function on $\mathbf{R}^n$. For the purposes of this work we are mainly interested in the case in which $\mathbf{X}$ is an infinite dimensional functional space. In particular we will focus on the case in which $\mathbf{X}$ is an affine space with vector space V so that in particular for all $x \in \mathbf{X}$ and all $v \in V$ we have that $x + v \in \mathbf{X}$. In this case then it is particularly straightforward to generalize the usual concept of directional derivative in the direction v at point x_0 as follows:

$$\delta F(x_0, v) := \lim_{s \to 0} \frac{F(x_0 + sv) - F(x_0)}{s}. \tag{1}$$

In general this quantity is called *Gâteaux differential*, or more traditionally *first variation*. The term Gâteaux differential comes from the notion of Gâteaux differentiability in Banach spaces (see [19]). Notice also that if we define $\psi(s) := F(x_0 + sv)$, then $\delta F(x_0, v) = \psi'(0)$.

This quantity is particularly important for the study of extremals of a functional since, as it happens for real valued functions, the vanishing of this quantity for all $v \in V$ it is a necessary condition to be satisfied by any local extremum of F.

Integral Functional and Euler equations. We will now restrict ourselves to functionals of the form

$$F(x) := \int_0^T L(t, x(t), \dot{x}(t), \ddot{x}(t))\, dt, \tag{2}$$

where $x \in \mathbf{X}$ and $L(t, z, p, a)$ is a continuous real valued function of the variables (t, z, p, a). Suppose furthermore that for example* $\mathbf{X} \subset C^2([0, T]; \mathbf{R}^n)$.

Now we can use the following well known result about integration that essentially says that under appropriate regularity assumptions *the derivative of the*

* A milder assumption would be $\mathbf{X} \subset H^2((0, T); \mathbf{R}^n)$.

integral is the integral of the derivative. More precisely if $f\colon [0,T]\times[-\tau,\tau]\to\mathbf{R}$, then if we let $\psi(s) := \int_0^T f(t,s)\,dt$ we have that

1. If f is continuous in $[0,T]\times[-\tau,\tau]$ then ψ is continuous in $[-\tau,\tau]$;
2. If f_s is continuous in $[0,T]\times[-\tau,\tau]$ then ψ' exists and it is given by

$$\psi'(s) := \int_0^T f_s(x,s)\,dx. \tag{3}$$

If we take, as we remarked above $\psi(s) := F(x+sv)$, then

$$\psi'(s) = \int_0^T \frac{d}{ds} L(t, x+sv, \dot{x}+s\dot{v}, \ddot{x}+s\ddot{v}) \tag{4}$$

Then, using the chain rule on L we have that the first variation (1) looks like

$$\begin{aligned}\delta F(x,v) = \int_0^T \big(&L_z(t,x(t),\dot{x}(t),\ddot{x}(t))\cdot v(t) + L_p(t,x(t),\dot{x}(t),\ddot{x}(t)\cdot\dot{v}(t)\\ &+ L_a(t,x(t),\dot{x}(t),\ddot{x}(t))\cdot\ddot{v}(t)\big)\,dt\end{aligned} \tag{5}$$

Now consider the relation $\delta F(x,v)=0$ for all $v\in C^\infty$; here we are considering $v\in C^\infty$ instead of C^2 functions since this has the advantage that we can consider the same class of variations for all differential equations of all order; moreover it is consistent with the usual conventions in the theory of distributions. This condition is equivalent to

$$\begin{aligned}\int_0^T \big(&L_z(t,x(t),\dot{x}(t),\ddot{x}(t))\cdot v(t) + L_p(t,x(t),\dot{x}(t),\ddot{x}(t)\cdot\dot{v}(t)\\ &+ L_a(t,x(t),\dot{x}(t),\ddot{x}(t))\cdot\ddot{v}(t)\big)\,dt = 0 \quad \forall v\in C^\infty((0,T);\mathbf{R}^n)\end{aligned} \tag{6}$$

This condition it is usually called the weak Euler equation for x. Notice that in order for this condition to be well defined the function $x(t)$ does not need more regularity than that declared in the definition of $\mathbf{X}$.

Now suppose that we take v vanishing at the boundary, then by integration by parts we get

$$\int_0^T \big(L_z(t,x(t),\dot{x}(t),\ddot{x}(t)) - \frac{d}{dt}L_p(t,x(t),\dot{x}(t),\ddot{x}(t)) + \frac{d^2}{dt^2}L_a(t,x(t),\dot{x}(t),\ddot{x}(t))\big)\cdot v(t)\,dt = 0. \tag{7}$$

This, of course, can be done if we have enough regularity on both L and x. This integral relation can be turned into a differential equation by using the fundamental lemma of calculus of variations:

Lemma. *Let $f\colon[0,T]\to\mathbf{R}^n$ be continuous function. If*

$$\int_0^T f(t)\cdot v(t)\,dt = 0 \quad \forall v\in C_c^\infty((0,T);\mathbf{R}^n), \tag{8}$$

then $f\equiv 0$ in $[0,T]$.

We therefore have

$$L_z(t, x(t), \dot{x}(t), \ddot{x}(t)) - \frac{d}{dt} L_p(t, x(t), \dot{x}(t), \ddot{x}(t)) + \frac{d^2}{dt^2} L_a(t, x(t), \dot{x}(t), \ddot{x}(t)) = 0. \tag{9}$$

Till now we have not taken into account the cases in which the variations does not vanishes at the boundary; this for example happens when if we formulate a variational problem without specifying in $\mathbf{X}$ the value of the solution at the boundaries.

Notice however that since the fundamental lemma of the calculus of variation the vanishing condition (8) need just to be verified for compactly supported functions, if we can prove that the solution of the variational problem is regular enough then the differential Eq. (9) will hold regardless of any boundary conditions. This being said let us now see what happens to the stationarity condition $\delta F(x, v) = 0$ when we do not assume the vanishing of the variation at the boundaries.

We have already discussed the fact that Eq. (9) still holds; therefore we are left with the contributions only from the boundary terms of the integration by parts; namely

$$\left[\left(L_p - \frac{d}{dt} L_a \right) v \right]_0^T + [L_a \dot{v}]_0^T = 0. \tag{10}$$

In order for this term to be zero the only possibility, other than the vanishing of v is to have

$$L_p - \frac{d}{dt} L_a = 0 \quad \text{and} L_a = 0, \tag{11}$$

at the boundary where we do not know that the variation is vanishing. This kind of boundary conditions usually are referred to as Neumann boundary conditions since they generally depends on the derivatives of the solution.

APPENDIX B

GAUSSIAN GREEN FUNCTIONS

And now there came both mist and snow, And it grew wondrous cold:
And ice, mast-high, came floating by,
As green as emerald.

— SAMUEL TAYLOR COLERIDGE, *The Rime of the Ancient Mariner*, (I, st. 13)

FIRST OF ALL define $L^m_\sigma := \sum_{n=0}^m (-1)^n (\sigma^{2n}/2^n n!) d^{2n}/dx^{2n}$ and consider the induced function $\rho_\sigma(x) := L^m_\sigma G_\sigma(x)$. Now notice that as $\sigma \to 0$, $L \to 1$ and $G_\sigma \to \delta$ so that we automatically have $LG = \delta$. It is also immediate to see that $\int \rho_\sigma(x)\, dx = 1$. Let us now consider for any $\delta > 0$

$$\lim_{\sigma\to 0} \int_{|x|\geq\delta} |\rho_\sigma(x)|\, dx = \lim_{\sigma\to 0} \sum_{n=1}^m \frac{\sigma^{2n}}{2^n n!} \int_{|x|\geq\delta} \left| \frac{d^{2n}}{dx^{2n}} G_\sigma(x) \right| dx;$$

notice that the sum starts at $n = 1$ because, since the gaussian is a mollifier, we already known that $\lim_{\sigma\to 0} \int_{|x|\geq\delta} e^{-x^2/2\sigma^2}/\sqrt{2\pi\sigma^2}\, dx = 0$. Now, let us consider the Hermite polynomial:

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n},$$

which satisfies

$$\frac{d^n G_\sigma(x)}{dx^n} = (-1)^n \frac{H_n(x/\sqrt{2}\sigma)}{(\sqrt{2}\sigma)^n} G_\sigma(x).$$

If we change the variable $y = x/\sigma$, we have

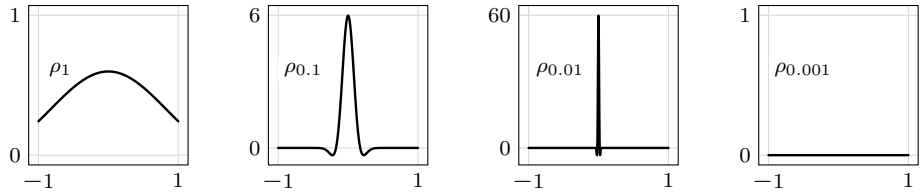
$$\begin{aligned} \sigma^{2n} \int_{|x|\geq\delta} \left| \frac{d^{2n}}{dx^{2n}} G_\sigma(x) \right| dx &= \frac{1}{2^n} \int_{|x|\geq\delta} \left| H_{2n}(x/\sqrt{2}\sigma) G_\sigma(x) \right| dx \\ &= \frac{1}{2^n} \int_{|y|\geq\delta/\sigma} \left| H_{2n}(y/\sqrt{2}) G_1(y) \right| dy, \end{aligned}$$

that goes to 0 as $\sigma \to 0$

To sum up, we can state the following lemma for ρ_σ.

Lemma. *The family of functions $\{\rho_\sigma\}_{\sigma>0} \subset C^\infty(\mathbf{R})$ has the following properties*

i. $\sup_{x\in\mathbf{R}} |\rho_\sigma(x)| < \infty$;
ii. $\int_{\mathbf{R}} \rho_\sigma(x)\, dx = 1$;



(a) (b) (c) (d)

Fig. 15. Plots of ρ_σ when $m = 1$ for various values of σ: $\sigma = 1$ in (a), $\sigma = 0.1$ in (b), $\sigma = 0.01$ in (c) and $\sigma = 0.001$ in (d).

iii. $\lim_{\sigma\to 0} \int_{|x|\geq\delta} |\rho_\sigma(x)|\,dx = 0$ *for every* $\delta > 0$.
iv. $\int_{\mathbf{R}} |\rho_\sigma(x)|\,dx < \infty$

In view of this Lemma the following theorem holds:

Theorem. *The sequence* $\langle \rho_\sigma \rangle$ *converges in* $\mathcal{D}'$ *to the* δ *function as* $\sigma \to 0$*:*

$$\lim_{\sigma\to 0} \int \rho_\sigma(x)\varphi(x)\,dx = \varphi(0), \qquad \forall \varphi \in C_c^\infty(\mathbf{R}).$$

Proof. Because of *ii.* we set

$$I_\sigma := \int \rho_\sigma(x)\varphi(x)\,dx - \varphi(0) = \int \rho_\sigma(x)\big(\varphi(x) - \varphi(0)\big)\,dx,$$

so that for any fixed $\delta > 0$:

$$I_\sigma = \int_{|x|\leq\delta} \rho_\sigma(x)\big(\varphi(x) - \varphi(0)\big)\,dx + \int_{|x|>\delta} \rho_\sigma(x)\big(\varphi(x) - \varphi(0)\big)\,dx;$$

if we let α_σ^δ be the first integral, and β_σ^δ the second one, we have the following bounds

$$|\alpha_\sigma^\delta| \leq \sup_{|x|\leq\delta} |\varphi(x) - \varphi(0)| \int_{|x|\leq\delta} |\rho_\sigma(x)|\,dx \leq K \sup_{|x|\leq\delta} |\varphi(x) - \varphi(0)| \equiv A_\delta,$$

$$|\beta_\sigma^\delta| \leq \sup_{|x|>\delta} |\varphi(x) - \varphi(0)| \int_{|x|>\delta} |\rho_\sigma(x)|\,dx \leq 2\|\varphi\|_{L^\infty(\mathbf{R})} \int_{|x|>\delta} |\rho_\sigma(x)|\,dx.$$

Now because of the continuity in $x = 0$ of φ and because of property *iii.* of the above lemma we have that $\lim_{\delta\to 0} A_\delta = 0$, and $\lim_{\sigma\to 0} \beta_\sigma^\delta = 0$ for every δ positive. Then

$$\max_{\sigma\to 0} \lim |I_\sigma| \leq A_\delta + \max_{\sigma\to 0} \lim |\beta_\sigma^\delta| = A_\delta,$$

now taking the limit $\delta \to 0$ we finally obtain $\lim_\sigma \to 0|I_\sigma| = 0$. ∎

APPENDIX C

PUBLICATIONS

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover all the tracks, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on.

— RICHARD P. FEYNMAN, Nobel Lecture (1965)

HERE IS a list of publications and reports that I produced during my PhD course.

Journal Papers.

1. **Alessandro Betti**, Stefano Bolognesi, Sven Bjarke Gudnason, Kenichi Konishi and Keisuke Ohashi, "Large-N $\mathbf{CP}^{N-1}$ sigma model on afinite interval and the renormalized string energy", *Journal of High Energy Physics*, 2018. **Candidate's contributions:** Carried out theoretical analyses.
2. **Alessandro Betti**, Marco Gori, Stefano Melacci "Cognitive Action Laws: The Case of Visual Features", *IEEE transactions on neural networks and learning systems*, 2019. **Candidate's contributions:** Carried out theoretical analyses, algorithm design and part of the experimental campaign.

Peer reviewed conference papers.

1. **Alessandro Betti**, Marco Gori, Giuseppe Marra "A Constrained-Based Approach to Machine Learning", *14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2018. **Candidate's contributions:** Carried out theoretical analyses
2. **Alessandro Betti**, Marco Gori and Stefano Melacci) "Motion Invariance in Visual Environments", *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. **Candidate's contributions:** Carried out theoretical analyses, algorithm design and part of the experimental campaign.
3. **Alessandro Betti**, Marco Gori "Least Action Principles and Well-Posed Learning Problems", To appear in the AIRO Springer Series volume associated with the ODS 2019 Conference. **Candidate's contributions:** Carried out theoretical analyses.
4. **Alessandro Betti**, Marco Gori "Trees in the Real Field", In: Esposito A., Faundez-Zanuy M., Morabito F., Pasero E. (eds) Neural Approaches to Dynamics of Signal Exchanges. Smart Innovation, Systems and Technologies, vol 151. Springer, Singapore. **Candidate's contributions:** Carried out theoretical analyses.

5. Giuseppe Marra, Dario Zanca, **Alessandro Betti**, and Marco Gori "Learning Activation Functions by means of Kernel Based Neural Networks". To appear in the AIIA conference proceedings, 2019. **Candidate's contributions:** Carried out theoretical analyses and part of the algorithms design.

Papers under review.

1. **Alessandro Betti**, Marco Gori and Stefano Melacci "Learning Visual Features Under Motion Invariance", currently submitted to Neural Networks. **Candidate's contributions:** Carried out theoretical analyses, algorithm design and part of the experimental campaign.

Other Publications (unrefereed contributions)**.**

1. **Alessandro Betti**, Marco Gori "Convolutional Networks in Visual Environments", `arxiv: cs.CV/1801.07110`, 2018. **Candidate's contributions:** Carried out theoretical analyses.
2. Giovanni Bellettini, **Alessandro Betti**, and Marco Gori "Generalization in quasi-periodic environments", `arXiv: cs.LG/1807.05343`, 2018. **Candidate's contributions:** Carried out theoretical analyses.
3. **Alessandro Betti**, Marco Gori, Giuseppe Marra "Backpropagation and Biological Plausibility", `arXiv: cs.AI/1808.06934`, 2018. **Candidate's contributions:** Carried out theoretical analyses.
4. **Alessandro Betti**, Marco Gori "On the Role of Time in Learning", `arXiv: cs.LG/1907.06198`, 2019. **Candidate's contributions:** Carried out theoretical analyses.
5. **Alessandro Betti**, Marco Gori, "Spatiotemporal Local Propagation", `arXiv: cs.LG/1907.05106`, 2019. **Candidate's contributions:** Carried out theoretical analyses, algorithm design and conducted experiments.
6. **Alessandro Betti**, Marco Gori, Marco Maggini, Giuseppe Marra, Stefano Melacci and Matteo Tiezzi, "Local Propagation in Constraint-based Neural Networks", Technical Report, 2019. **Candidate's contributions:** Carried out theoretical analyses.

APPENDIX D

BIBLIOGRAPHY

You will find it a very good practice always to verify your references, sir.

— MARTIN JOSEPH ROUTH, from J. W. Burgon, *Memoir of Dr. Routh* (1878)

HERE ARE THE REFERENCES to the works cited in the various chapters. They are written following the alphabetical order of the authors.

1 Milton Abramowitz, Irene Ann Stegun, *Handbook of Mathematical Functions*, United States Department of Commerce, National Bureau of Standards, ninth printing 1970.

2 Andrillon T., N. Yuval N., C. Cirelli, G. Tononi and F. Itzhak, "Single-neuron activity and eye movements during human REM sleep and awake vision" *Nature Communications* **6** (2015), 7884.

3 Vladimir I. Arnold, *Metodi matematici della meccanica classica.* Editori riuniti university press (2010).

4 Simon Baker, Daniel Scharstein, J.P. Lewis, Stefan Roth, Michael J. Black, Richard Szeliski, "A Database and Evaluation Methodology for Optical Flow," *Int. J. Comput. Vision* **92** (2011), 1–31.

5 Alessandro Betti, Marco Gori, "The principle of least cognitive action," *Theoretical Computer Science* **633** (2016), 83–99.

6 Alessandro Betti, Marco Gori, Giuseppe Marra, "Backpropagation and Biological Plausibility," `arXiv:cs.AI/1808.06934` (2018).

7 Alessandro Betti, Marco Gori, Stefano Melacci, "Cognitive Action Laws: The Case of Visual Features." *IEEE transactions on neural networks and learning systems* (2019).

8 Ludwig Boltzmann, "Studien über das gleichgewicht der lebendigen kraft zwischen bewegten materiellen punkten," *Wiener Berichte* **58** (1866), 49–96.

9 Léon Bottou, Frank E. Curtis, and Jorge Nocedal, "Optimization Methods for Large-Scale Machine Learning," *SIAM Rev.* **60** (2018), 223–311.

11 Braddick O. and J. Atkinson, "Development of human visual function," *Vision Research* **51** 2011, 1588–1609.

10 Haim Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer, 2010.

12 Roger W. Brockett, *Finite Dimensional Linear Systems.* SIAM Edition, 1970.

13 Miguel Á . Carreira-Perpiñán, Weiran Wang, "Distributed Optimization of Deeply Nested Systems," *Artificial Intelligence and Statistics* (2014), 10–19.

14 Ennio De Giorigi. "Congetture riguardanti alcuni problemi di evoluzione," *Duke Math. J.* **80** (1995), 255–268.

15 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE conference on computer vision and pattern recognition,* (2009), 248–255.

16 Charles A. Desoer and Mathukumalli Vidyasagar, *Feedback Systems: Input-Output Properties.* New York: Academic Press, 1975

17 Dobson V. and D. Y. Teller, "Visual acuity in human infants: A review and comparison of behavioral and electrophysiological studies," *Vision Research* **18** (1978), 1469-83.

18 Giovanni Gallavotti, *The elements of mechanics.* Springer Science & Business Media, 2013.

19 Mariano Giaquinta, Stefan Hildebrant, *Calculus of Variations.* Springer, 2004.

20 Melvyn A. Goodale and A. David Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences* **1** (1992), 20–25.

21 Marco Gori, *Machine Learning: A Constrained-Based Approach.* Morgan Kaufmann, 2017.

22 Marco Gori, Marco Lippi, Marco Maggini, Stefano Melacci, "Semantic video labeling by developmental visual agents," *Computer Vision and Image Understanding* **146** (2016), 9–26.

23 Marco Gori, Marco Maggini, Alessandro Rossi, "Neural network training as a dissipative process," *Neural Networks* **81** (2016), 71–80.

24 Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, Yann LeCun, "Unsupervised learning of spatiotemporally coherent metrics," *Proceedings of the IEEE International Conference on Computer Vision* (2015), 4086–4093.

25 Akhilesh Gotmare, Valentin Thomas, Johanni Brea, Martin Jaggi, "Decoupling Backpropagation using Constrained Optimization Methods," *Workshop on Efficient Credit Assignment in Deep Learning and Deep Reinforement Learning, ICML 2018* (2018), 1–11.

26 Jack K. Hale, *Ordinary Differential Equations.* Krieger Publishing Company Malabar, Florida, 1969.

27 L. Herrera, L. Nunez, A. Patino, H. Rago, "A variational principle and the classical and quantum mechanics of the damped harmonic oscillator," *American Journal of Physics* **54** (1986), 273–277.

28 Berthold K.P. Horn and Brian G. Schunck, "Determining optical flow," *Artificial Intelligence* **17** (1981), 185–203.

29 Kerson Huang, *Statistical Mechanics.* John Wiley & Sons; second edition, 1963.

30 D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex," *J. Physiol.* **160** (1962) 106154.

31 E.T. Jaynes, "Information theory and statistical mechanics," *The Phisical Review* **106** (1957), 620–630.

32 Koray Kavukcuoglu, Pierre Sermanet, Y-lan Boureau, Karol Gregor, Michael Mathieu, Yann L. Cun, "Learning Convolutional Feature Hierarchies for Visual Recognition," *Advances in neural information processing systems* (2010), 1090–1098.

33 L. Landau et E. Lifchitz, *Physique Théorique Tome 1: Mécanique.* Éditions Mir (1966).

34 Yann LeCun, "A theoretical framework for back-propagation," *Proceedings of the 1988 connectionist models summer school. Vol. 1. CMU, Pittsburgh, Pa: Morgan Kaufmann*, 1988.

35 Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," *Proceedings of the 26th Annual International Conference on Machine Learning* (2009).

36 Yin Li, Manohar Paluri, James M. Rehg and Piotr Dollár, "Unsupervised Learning of Edges," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 1619–1627.

37 Matthias Liero, Ulisse Stefanelli, "A New Minimum Principle for Lagrangian Mechanics," *J. Nonlinear Sci.* **23** (2013), 179–204.

38 David G. Lowe,"Distinctive Image Features from Scale-Invariant Keypoints," *International journal of computer vision* **60** (2006) 91–110.

39 David Marr, *Vision.* Freeman, San Francisco, 1982.

40 Stefano Melacci and Marco Gori, "Unsupervised Learning by Minimal Entropy Encoding," *IEEE Trans. Neural Netw. Learning Syst.* **23** (2012), 1849–1861.

41 James L. McClelland, David E. Rumelhart, and PDP Research Group, *Parallel Distributed Processing, Volume 1.* MIT press, 1987.

42 William of Occam, *Quaestiones et decisiones in IV libros Sententiarum, cum centilogio theologico.* Johann Trechsel.

43 Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, Bharath Hariharan, "Learning Features by Watching Objects Move" *CVPR 2017.*

44 Tommaso Poggio, Federico Girosi, "A theory of networks for approximation and learning," No. AI-M-1140. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB (1989).

45 Tommaso Poggio, Fabio Anselmi, *Visual Cortex and Deep Networks: Learning Invariant Representations.* MIT Press, 2016.

46 Boris T. Polyak, "Some methods of speeding up the convergence of iteration methods," *U.S.S.R. Comput. Math. Math. Phys.* **4** (1964), 1–17.

47 Marc' Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau, Yann LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), 1–8.

48 Carlo Rovelli, "Aristotle's Physics: A Physicist's Look," *Journal of the American Philosophical Association* **1** (2015), 23–40.

49 Enrico Serra, Paolo Tilli, "Nonlinear wave equations as limits of convex minimization problems: proof of a conjecture by De Giorgi," *Annals of Mathematics* **175** (2012), 1551–1574.

50 Ulisse Stefanelli, "The De Giorgi conjecture on elliptic regularization," *Math. Models Methods Appl. Sci.* **21** (2011), 1377–1394.

51 Sun, J., Wang, Q., and Zhong, Q., "A less conservative stability test for second-order linear time-varying vector differential equations," *Int. J. Control* **80** (2007), 523–526.

52 Lin Sun, Kui Jia, Tsung-Han Chan, Yuqiang Fang, Gang Wang, Shuicheng Yan, "DL-SFA: deeply-learned slow feature analysis for action recognition," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2014), 2625–2632.

53 Amirhossein Tavanaei, Timothée Masquelier and Anthony S. Maida, "Acquisition of Visual Features Through Probabilistic Spike-Timing-Dependent Plasticity," *2016 International Joint Conference on Neural Networks* (2016).

55 Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, Tom Goldstein, "Training Neural Networks Without Gradients: A Scalable ADMM Approach," *International conference on machine learning* (2016), 2722–2731.

54 Lorenzo Tentarelli, Paolo Tilli "An existence result for dissipative nonhomogeneous hyperbolic equations via a minimization approach," *Journal of Differential Equations* **2668** (2019), 5185–5208.

56 Vladimir N. Vapnik, *Statistical Learning Theory.* John Wiley & Sons, 1998.

57 Peter Walter, *An introduction to Ergodic Theory.* Vol. 79. Springer Science & Business Media, 2000.

58 Xiaolong Wang, Abhinav Gupta, "Unsupervised Learning of Visual Representations Using Videos," *The IEEE International Conference on Computer Vision* (2015).

59 Steven Weinberg, *The Quantum Theory of Fields* Vol. 2. Cambridge university press, 1995.

60 Laurenz Wiskott and Terrence J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation* **14** (2002), 715–770.

61 Laurenz Wiskott, "Slow feature analysis: A theoretical analysis of optimal free responses," *Neural Computation* **15** (2003), 2147–2177.

62 Dario Zanca and Marco Gori, "Variational Laws of Visual Attention for Dynamic Scenes," *Advances in Neural Information Processing Systems 30* (2017), 3826–3835.

63 Dario Zanca, Stefano Melacci, Marco Gori, "Gravitational Laws of Focus of Attention," *IEEE transactions on pattern analysis and machine intelligence* (2019).

64 C. Lawrence Zitnick, Piotr Dollár, "Edge Boxes: Locating Object Proposals from Edges," *Computer Vision–ECCV 2014* (2014).