Variational Foundations of Online Backpropagation

S. Frandina, M. Gori, M. Lippi, M. Maggini, S. Melacci

Department of Information Engineering and Mathematics
University of Siena

Abstract. On-line Backpropagation has become very popular and it has been the subject of in-depth theoretical analyses and massive experimentation. Yet, after almost three decades from its publication, it is still surprisingly the source of tough theoretical questions and of experimental results that are somewhat shrouded in mystery. Although seriously plagued by local minima, the batch-mode version of the algorithm is clearly posed as an optimization problem while, in spite of its effectiveness in many real-world problems, the on-line mode version has not been given a clean formulation, yet. Using variational arguments, in this paper, the on-line formulation is proposed as the minimization of a classic functional that is inspired by the principle of minimal action in analytic mechanics. The proposed approach clashes sharply with common interpretations of on-line learning as an approximation of batch-mode, and it suggests that processing data all at once might be just an artificial formulation of learning that is hopeless in difficult real-world problems.

Keywords: on-line Backpropagation, principle of least action, regularization, local minima, dissipative systems.

1 Introduction

In classical statistics, sum-minimization problems arise in least squares and in maximum-likelihood estimation (for independent observations). The general class of estimators that arise as minimizers of sums are called M-estimators. Backpropagation [7] was proposed to efficiently compute the gradient of the cost function associated with a supervised neural network. In spite of the plain numerical computation of the gradient, in many cases, it makes it possible to break the barrier that enables many application of neural networks to real-world problems. Unfortunately, the convergence of the algorithm is seriously plagued by the presence of local minima in the error function [4]. In many cases, instead of performing a classic gradient descent scheme, the gradient computation for single examples (on-line mode) has been profitably used by updating directly the parameters, without accumulating those contributions for all the training set. The on-line scheme is especially adequate to real-world problems where the examples are streamed continuously in time. There is plenty of evidence that such a stochastic gradient descent has been very effective in the case of large-scale

Links with Analytic Mechanics		
variable	machine learning	analytic mechanics
w_i	weight	particle position
$\dot{w}_i$	weight variation	particle velocity
V	loss temporal derivative	potential energy
T	temporal smoothness	kinetic energy
$\mathcal{L} = T - V$	Cognitive Lagrangian	Mechanical Lagrangian
$\mathcal{S} = \int_0^{t_e} \mathcal{L}\ dt$	Cognitive Action	Mechanical Action

Table 1. Links between machine learning and analytic mechanics.

problems [1]. Amongst others, the Backpropagation on-line training scheme is often regarded as a way to get around shallow local minima of the cost function, but like for the batch-mode scheme, it is quite hard to understand the conditions of convergence, apart from relative simple cases [3].

After almost three decades from its publication, on-line Backpropagation is still surprisingly the source of tough theoretical questions, and it has not received a fully satisfactory formulation, yet. Using variational arguments, in this paper, the on-line formulation is proposed as the minimization of a classic functional that is inspired by the principle of least action in analytic mechanics. However, the classic Lagrangian is replaced with a *time-variant* function that is responsible of a dissipative behavior that plays a major role in any learning process. We prove that a "strong dissipation" transforms the continuous time differential law coming from the Euler-Lagrange equation into the classic on-line Backpropagation with its stochastic gradient numerical computation. The proposed approach clashes sharply with common interpretations of on-line learning as an approximation of batch-mode. On the other hand, differently from what is generally assumed, it suggests that processing data all at once might be just an artificial formulation of learning that is hopeless in difficult real-world problems.

2 On-line Backpropagation revisited

We consider a feedforward neural network as a function which transforms a given input $x \in \mathbb{R}^d$ into a real number, that is $f : (x, w) \in \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, being w the vector of weights and x the input. The analysis carried out in this paper does not make any hypothesis on the network structure and, consequently, on $f(x, w)$, but we like to think of it in terms of feedforward networks mostly because of their universal approximation capabilities and their biological inspiration [6].

POTENTIAL ENERGY

Now we introduce the loss function $\overline{V}(f, y)$, along with the set of supervised pairs $\mathcal{P} = \{(x_\kappa, y_\kappa)\}_{\kappa=1}^{\ell}$. For example, $\overline{V}(f, y)$ can be the hinge function – typical for classification – or the quadratic function $(f(x_\kappa) - y_\kappa)^2$ – typical of regression.

Let ζ be a mollifier. As an example, we can choose

$$\zeta_\epsilon(\tau) = \begin{cases} Z_\epsilon \cdot \exp\left(1 - \epsilon^2/(\epsilon^2 - \tau^2)\right) & \text{if } |\tau| < \epsilon \\ 0 \text{ if } |\tau| \geq \epsilon, \end{cases}$$

where Z_ϵ is taken in such a way that $\int_{-\infty}^{+\infty} \zeta(\tau) d\tau = 1$. A nice property of mollifiers is their weak convergence to the Dirac distribution, that is

$$\lim_{\epsilon \to 0} \zeta(\tau) = \delta(\tau).$$

Let $[0, t_e]$ be the time interval, $t_0 = 0$ with $t_e > 0$[1]. Now, let t_κ be the time instant at which the pair (x_k, y_k) becomes available, let $t_\kappa < t_e$ be, and consider the functional

$$\mathcal{V}(w) = \int_0^{t_e} \psi(t) V(w(t)) dt$$

where

$$V(w(t)) = \sum_{\kappa=1}^{\ell} \zeta(t - t_\kappa) \cdot \overline{V}(f(x(t), w(t)), y(t))).$$

and $\psi \in C^\infty([0, t_e], I\!R^+)$ is a monotone increasing function which, in this paper, is chosen as $\psi(t) = e^{\beta t}$. As it will be shown later, this is related to energy dissipation and plays a crucial role for the establishment of the learning process. Basically, it prescribes a growing weight of the loss as time evolves. Now, let us assume $w \in I\!R^m$. We start to regard it as the *Lagrangian coordinates* of a virtual mechanical system. The learning problem defined by the supervised pairs $\mathcal{P}$, for a given choice of weights $w(t)$ at time t, defines a function $V(w(t))$ that, throughout this paper, is referred to as the *potential energy* of the system (neural network) defined by Lagrangian coordinates w. In machine learning, we look for trajectories $w(t)$ that possibly lead to configurations with small potential energy. The classic supervised learning is given a more adequate interpretation as $\epsilon \to 0$, which leads to replace the mollifiers with correspondent Dirac distributions $\delta(t - t_\kappa)$. When choosing the quadratic loss, we get

$$V(w(t)) = \sum_{\kappa=1}^{\ell} \delta(t - t_\kappa) \cdot (y(t) - f(x(t), w(t)))^2 .$$

The learning process in this case is only expected to reduce the error corresponding to the supervision points. For binary classification problems with $y(t) \in \{-1, 1\}$, however, if we adopt the hinge function

$$V(w(t)) = \sum_{\kappa=1}^{\ell} \delta(t - t_\kappa) \cdot \max\{\gamma - y(t) \cdot f(x(t), w(t)), 0\}$$

[1] It is of interest to consider also the case in which $t_e = \infty$.

being $\gamma > 0$ a proper threshold, we can promptly see that the learning process can led to the perfect match (zero loss) on some of the examples of the training set.

Now, following the duality with mechanics, we introduce the kinetic energy.

KINETIC ENERGY
Let $\mu_i > 0$ be the *mass* of each particle defined by the *position* $w_i(t)$ and *velocity* $\dot{w}_i$. Then, let us consider the *kinetic energy*

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

It gives a glimpse of the converge of the process of learning, since its end corresponds with $T = 0$. Like for the potential energy, in this paper we are interested in the accumulation $\int_0^{t_e} e^{\beta t} T(t) dt$ over $[0, t_e]$, which reflects the smoothness of the velocities of the particles. Moreover, also for the kinetic energy, we provide a growing account as time evolves which, as already stated, will be shown to be the basis of a dissipative behavior. The introduction of the exponential factor in both the potential and kinetic energy has been proposed in analytic mechanics as a way of introducing dissipation processes that are not present within the pure Hamiltonian framework [5].

VARIATIONAL FORMULATION OF LEARNING
Let us introduce the Lagrangian

$$\mathcal{L} := T - V$$

The problem of online learning can be formulated as that of finding

$$w^\star = \arg\min_{w \in \mathcal{W}} \int_0^{t_e} e^{\beta t} \mathcal{L}(w(t)) dt \tag{2}$$

3 Backprop from Euler-Lagrange equations

The solution of the online learning problem can be obtained by finding stationary points of (2).

Theorem 1. *The solution of online learning stated by (2) satisfies*

$$\ddot{w}_i^\star + \beta \dot{w}_i^\star + \frac{1}{\mu_i} V'_{w_i} = 0, \tag{3}$$

where $V'_{w_i} = \sum_{\kappa=1}^{\ell} \overline{V}'_{w_i} \, \delta(t - t_\kappa)$.

Proof. We have

$$\frac{d}{dt} \frac{\partial}{\partial \dot{w}_i} \left(e^{\beta t} \mathcal{L} \right) = \frac{d}{dt} \frac{\partial}{\partial \dot{w}_i} \left(e^{\beta t} T \right) = \mu_i \frac{d}{dt} \left(e^{\beta t} \dot{w}_i \right) = \mu_i \left(e^{\beta t} \ddot{w}_i + \beta e^{\beta t} \dot{w}_i \right)$$

and

$$\frac{\partial}{\partial w_i}\left(e^{\beta t}\mathcal{L}\right) = -e^{\beta t}\frac{\partial V}{\partial w_i} = -e^{\beta t}V'_{w_i}.$$

Then the thesis follows when applying the Euler-Lagrange equation of (2). QED.

Notice that, since this theorem comes from the Euler-Lagrange equations, like for the action in analytic mechanics, the solution of the equations is not necessarily the absolute minimum. In general, it is a stationary point which is typically a saddle point. Interestingly, as shown in Section 4, like for other physical laws, this stationary point has nice minimization properties on the potential energy, which is exactly what we look for also in learning. Now let us assume that the system evolve from null Cauchy's conditions $w_i(0) = \dot{w}_i(0) = 0$ and let us use the notation $g_{i,\kappa} := \overline{V}'_{w_i}(w_i(t_\kappa))$. The following theorem holds true

Theorem 2. *The evolution from* null Cauchy's condition *follows the differential equation*

$$\frac{dw_i^\star}{dt} + \beta w_i^\star = -\frac{1}{\mu_i}\sum_{\kappa=1}^{\ell} g_{i,\kappa}\cdot 1(t-t_\kappa). \tag{4}$$

Proof. From Theorem 1 we have

$$\begin{aligned}\int_0^t \frac{d}{d\theta}\left(\frac{dw_i^\star}{d\theta} + \beta w_i^\star\right)\, d\theta &= -\frac{1}{\mu_i}\sum_{\kappa=1}^{\ell}\int_0^t \overline{V}'_{w_i}\cdot\delta(\theta-t_\kappa)\, d\theta \\ &= -\frac{1}{\mu_i}\sum_{\kappa=1}^{\ell} g_{i,\kappa}\cdot 1(t-t_\kappa).\end{aligned}$$

Now, the thesis follows when considering that $w_i^\star(0) = 0$ and $\dot{w}_i^\star(0) = 0$. QED.

Now, let us consider the answer to the first stimulus (supervised pair) coming at $t = t_1$ from the initial conditions $w_i(0) = \dot{w}_i(0) = 0$. We have

$$\frac{dw_i^\star}{dt} + \beta w_i^\star = -\frac{g_{i,1}}{\mu_i}.$$

If $w_i(0) = 0$ then

$$w_i^\star(t) = \frac{-g_{i,1}}{\beta\mu_i}\left(1 - e^{-\beta(t-t_1)}\right),$$

which indicates an asymptotic evolution to

$$\overline{w}_i^\star = \lim_{t\to\infty} w_i^\star(t) = \frac{-g_{i,1}}{\beta\mu_i}$$

Now we have $|w_i^\star(5/\beta) - \overline{w}_i^\star|/|\overline{w}_i^\star| < 0.01$, which means that with large values of β – or equivalently, small time constant $1/\beta$ – the weights are updated[2] from $w_i^\star(0) = w_i^\star|_0 = 0$ to $w_i^\star(1) = w_i^\star|_1$ by

$$w_i^\star(1) \simeq w_i^\star|_1 = -\eta_i \cdot g_{i,1} = -\frac{1}{\beta\mu_i} g_{i,1},$$

where $\eta_i := 1/(\beta\mu_i)$ is the classic *learning rate*. From now on, the notation $\simeq$ is used to indicate the above stated approximation of the asymptotic value $\overline{w}_i^\star$. Interestingly, the required high value for β corresponds with small learning rate, which is also kept small when considering particles with large mass μ_i. Beginning from this remark, now we establish the connection between the formulated continuous framework of learning with the classic on-line Backpropation algorithm.

Theorem 3. *Given $\mathcal{P} = \{x_\kappa, y_\kappa\}_{\kappa=1}^{\ell}$, where the supervised pairs (x_κ, y_κ) comes at $t = t_\kappa$, let us β such that $\forall \kappa = 1, \ldots, \ell$ we have*

$$\tau := 10/\beta \leq t_\kappa - t_{\kappa-1}. \tag{5}$$

Then

$$w_i^\star(t_k + \tau) \simeq w_i^\star(t_\kappa - \tau) - \eta_i g_{i,\kappa}, \tag{6}$$

which corresponds with the discrete counterpart

$$w_i^\star|_\kappa \simeq w_i^\star|_{\kappa-1} - \eta_i g_{i,\kappa}, \tag{7}$$

commonly referred to as the on-line Backpropagation algorithm.

Proof. We have

$$\int_{t_k-\tau}^{t_\kappa+\tau} \frac{d}{d\theta}\left(\frac{dw_i^\star}{d\theta} + \beta w_i^\star\right)\, d\theta = -\frac{1}{\mu_i}\sum_{\kappa=1}^{\ell}\int_{t_k-\tau}^{t_\kappa+\tau} \overline{V}'_{w_i}(w(t)) \cdot \delta(\theta - t_\kappa) d\theta,$$

from which we derive

$$\left(\frac{dw_i^\star}{d\theta} + \beta w_i^\star\right)_{t_\kappa+\tau} - \left(\frac{dw_i^\star}{d\theta} + \beta w_i^\star\right)_{t_\kappa-\tau} = -\frac{1}{\mu_i}\, 1(t - t_\kappa) g_{i,\kappa}.$$

Now, because of the strong damping hypothesis (5)

$$\left(\frac{dw_i^\star}{d\theta}\right)_{t_\kappa-\tau} \simeq 0 \;\; \textit{and} \;\; w_i^\star(t_\kappa - \tau) \simeq w_i^\star|_{\kappa-1}$$

and, therefore, for $t > t_\kappa$ we get

$$\frac{dw_i^\star}{d\theta}|_{t_{\kappa+\tau}} + \beta w_i^\star|_{t_{\kappa+\tau}} - \beta w_i^\star|_{\kappa-1} \simeq -\frac{1}{\mu_i} g_{i,\kappa}.$$

Finally, the thesis follows when invoking again the strong damping hypothesis (5).
QED.

[2] We use the notation $w_i^\star|_t$ to indicate the corresponding *discrete updating* that are used in the on-line Backpropagation algorithm.

4 Learning as a dissipative Hamiltonian process

Now we can establish a conservation principles that is related to dissipative systems[3]. From Theorem 1, if we multiply by $\dot{w}_i$ and accumulate over the weights, we get

$$\sum_{i=1}^{m} \mu_i \left(\dot{w}_i \ddot{w}_i + \beta \dot{w}_i^2 \right) + \sum_{\kappa=1}^{\ell} \sum_{i=1}^{m} \overline{V}'_{w_i}(w_i(t)) \ \dot{w}_i \ \delta(t - t_\kappa) = 0.$$

Now we have

$$\frac{d\overline{V}(w(\theta))}{d\theta} = \sum_{i=1}^{m} \overline{V}'_{w_i}(w_i(t)) \ \dot{w}_i.$$

If we accumulate over $[t_a, t_b]$ we get

$$\int_{t_a}^{t_b} \frac{d}{d\theta} \left(\frac{1}{2} \sum_{i=1}^{m} \mu_i \dot{w}_i^2 \right) \ d\theta + \int_{t_a}^{t_b} \frac{d\overline{V}(w(\theta))}{d\theta} \cdot \sum_{\kappa=1}^{\ell} \delta(\theta - t_\kappa) \ d\theta + \int_{t_a}^{t_b} \beta \sum_{i=1}^{m} \mu_i \dot{w}_i^2 \ d\theta = 0.$$

Now, if we define

$$D(t) := \int_0^t \beta \sum_{i=1}^{m} \mu_i \dot{w}_i^2 \ d\theta = \frac{1}{\eta_i} \sum_{i=1}^{m} \int_0^t \dot{w}_i^2 \ d\theta,$$

then, we get

$$\int_{t_a}^{t_b} \frac{d}{d\theta} \left(T + \overline{V} \sum_{\kappa=1}^{\ell} \delta(\theta - t_\kappa) + D \right) d\theta = 0.$$

Now, we use a notation overloading to denote by $T(t)$ the kinetic energy at t and we assume that $q \leq \ell$ supervised examples have been presented in $[0, t]$, begin $t \in [t_a, t_b]$. If $\nexists \kappa = 1, \ldots, \ell : \ t_\kappa \in [t_a, t_b]$ the above equation turns into the conservation equation

$$E(t) = T(t) + \overline{V} \sum_{\kappa=1}^{q} 1(t - t_\kappa) + D(t) = c \tag{8}$$

being c the constant energy of the extremes of the interval $[t_a, t_b]$. The overall energy $E(t)$ is conserved in all intervals in which there is no supervision. Whenever a supervised example is presented in the interval, the energy increases by

$$\overline{V} \sum_{\kappa=1}^{q} \left(1(t_\kappa) - 1(t_{\kappa-1}) \right).$$

It turns out that the energy is injected by any supervised pairs, which yield new potential energy that is partly transformed into kinetic energy and partly dissipated. It is in fact the strong dissipation hypothesis given in terms of β which is responsible of producing stochastic gradient descent and which ensures the convergence of the learning process.

[3] For the sake of simplicity, in the following we drop the symbol $\star$.

5 Conclusions

This paper gives a clean foundation of on-line Backpropagation in a variational framework with strong connections with analytic mechanics. This approach can be thought of as the temporal counterpart of the study on regularization in the feature space given in [2]. It is shown that learning is in fact a dissipative process and that if the damping parameter β is large enough then we end up in the classic stochastic gradient descent scheme of on-line Backpropagation. This formulation might be of interest for exploring the new frontiers of lifelong learning models, in which we abandon learning processes based on training sets and consider intelligent agents living in their own environments. To this purpose, we have given natural laws expressed by second-order differential equations that obey an intriguing principle of energy conservation.

While all this quenches the desire of giving Backpropagation a formulation that resembles that of classic laws of Nature, the most attracting picture emerges when forcing small values of β, namely small dissipation in the learning process. In so doing we depart significantly from stochastic gradient descent and our preliminary connections with Statistical Mechanics indicate that when learning with small dissipation we can gain more chance to get optimal solutions with respect to the traps of gradient descent. Intuitively, this is quite simple; the weights become particles whose behavior is that of a damping oscillation system, which is very well-suited to escape from local minima traps. Further research is needed to provide theoretical and experimental evidence of this intuition.

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

References

1. Bottou, L., Bousquet, O.: The tradeoffs of large-scale learning. Advances in Neural Information Processing Systems 20, 161–168 (2008)
2. Gnecco, G., Gori, M., Sanguineti, M.: Learning with boundary conditions. Neural computation 25(4), 1029–1106 (2013)
3. Gori, M., Maggini, M.: Optimal convergence of on-line backpropagation. Neural Networks, IEEE Transactions on 7(1), 251–254 (1996)
4. Gori, M., Tesi, A.: On the problem of local minima in backpropagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(1), 76–86 (1992)
5. Herrera, L., Nunez, L., Patino, A., Rago, H.: A variational principle and the classical and quantum mechanics of the damped harmonic oscillator. American Journal of Physics 53(3), 273 (1985)
6. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural networks 4(2), 251–257 (1991)
7. Rumelhart, D.E., Hintont, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (1986)