Extended Hamiltonian Learning on Riemannian Manifolds: Theoretical Aspects

Simone Fiori

***Abstract*—The present contribution introduces a general theory of extended Hamiltonian (second-order) learning on Riemannian manifolds, as an instance of learning by constrained criterion optimization. The dynamical learning equations are derived within the general framework of extended Hamiltonian stationary-action principle and are expressed in a coordinate-free fashion. A theoretical analysis is carried out in order to compare the features of the dynamical learning theory with the features exhibited by the gradient-based one. In particular, gradient-based learning is shown to be an instance of dynamical learning and the classical gradient-based learning modified by a 'momentum' term is shown to resemble discrete-time dynamical learning. moreover, the convergence features of gradient-based and dynamical learning are compared on a theoretical basis. The paper discusses cases of learning by dynamical systems on manifolds of interest in the scientific literature, namely, the Stiefel manifold, the special orthogonal group, the Grassmann manifold, the group of symmetric positive definite matrices, the generalized flag manifold and the real symplectic group of matrices.**

***Index Terms*—Extended Hamiltonian (second-order) learning; Riemannian manifold; Learning by constrained criterion optimization; Gradient-based (first-order) learning.**

I. Introduction

An instance of classical learning theory is by criterion optimization over an Euclidean space. Consider, as an illustrative example, the case of learning in a simple system formed by a single neuron. The connection weight-vector x belongs to the space $\mathbb{R}^n$, where n denotes the number of synapses, and the performance of the system on a given problem is measured by an error function $E : \mathbb{R}^n \to \mathbb{R}$ of the weight-vector, namely, $E(x)$. In classical learning theory, the optimal connection pattern is sought for in order to minimize the error E. The search for an optimal connection pattern may be effected by moving along a trajectory $x(t)$ within the search space $\mathbb{R}^n$, with $t \in \mathbb{R}$, following the direction of the Euclidean gradient of the error function, namely by [67]:

$$\frac{dx}{dt} = -\partial_x E, \tag{1}$$

where symbol ∂_x denotes the Euclidean gradient (or Jacobian) computed with respect to the weight-vector x. The implementation of the above learning theory on a computer is rather straightforward as the search space is flat, therefore, the differential equation (1) may be integrated numerically by means of one of the various stepping methods available in the literature, such as the Euler method or the Runge-Kutta method [28].

The learning theory (1) exhibits two distinguishing features: the search space is flat, namely, there are no constraints on the weight-vector values, and the learning equation only involves the velocity of the weight-vector and the gradient of the error function. A more advanced instance of adaptive system parameter learning is by constrained optimization. In such context, a (smooth) criterion function of the system parameters measures its learning performance and suitable constraints on parameters' values reflect the natural constraints presented by the learning problem. Whenever the constraints make the set of feasible learning parameters form a smooth manifold, learning takes place on differentiable manifolds. In this event, differential geometry is an appropriate mathematical instrument to formulate and to implement a learning theory. As an illustrative example of constrained optimization, consider the case of learning for a single neuron whose weight-vector is constrained by a energy conservation law, for example, $\|x\| = 1$, where symbol $\|\cdot\|$ denotes vector norm. Such problem arises whenever all that matters is the direction of the weight-vector x, while its norm and sign do not matter at all. For example, this is the case in principal component analysis [11], one-unit independent component analysis [38] and in neural blind deconvolution [17]. When the parameter space is not Euclidean but it has the structure of a curved Riemannian manifold M, the Euclidean-gradient-steepest-descent learning theory (1) should be replaced by a Riemannian-gradient-steepest-descent learning theory, that takes on the expression:

$$\frac{dx}{dt} = -\nabla_x E, \tag{2}$$

where $E : M \to \mathbb{R}$ and symbol ∇_x denotes the Riemannian gradient with respect to a chosen metrics. The equation (2) ensures that the learning trajectory $x(t)$ does not escape the feasible parameter space and hence does not break any learning constraints at any time. The idea conveyed by equation (2) may be traced back to the paper [37], although its numerical implementation appeared unpractical to the author of [37]. Recent advancements in the geometrical integration field made equation (2) quite popular in applied sciences [28]. Apparently, the learning theory (2) inherits those drawbacks that are well known about gradient-based optimization like, for instance, the slow-learning phenomenon in presence of plateaus in the error surface.

In order to generalize the learning theory (2), the present paper proposes a class of learning equations which stem from modern mechanics formalism and, in particular, from the the differential-geometrical formulation of mechanics [36], [42]. Indeed, a quite general framework to design a learning theory for constrained optimization is provided by the (extended) Hamiltonian formulation. Hamilton's stationary-action principle admits, as solution, a trajectory on the manifold of system parameters that maximizes system's learning performances and adjusts its kinetic energy accordingly. Moreover, the extended Hamiltonian principle may account for additional learning factors such as non-conservative forces which are reminiscent of the momentum term in classical learning theory. Through calculus of variations on manifolds, the minimum-action principle leads to a formulation of learning equations under the form of extended Hamiltonian (i.e., second-order) systems.

The theory of extended Hamiltonian systems proved fruitful in optimization and learning over the past years. An early contribution in literature that explored the benefits of extended Hamiltonian systems in (unconstrained) optimization is [2]. A thorough analysis of the convergence properties and of the improvement due to dynamical learning over gradient-based learning was conducted in [52], which

The author is with Dipartimento di Ingegneria Biomedica, Elettronica e Telecomunicazioni (DiBET), Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy. (eMail: s.fiori@univpm.it)

refers to the only case of dynamical learning over the Euclidean space $\mathbb{R}^n$. More recently, contributions have appeared in the neural networks and signal processing literature that explored the profitability of extended Hamiltonian systems in specific applications [14], [15].

The extended Hamiltonian learning equations are formulated in a coordinate-free fashion. The coordinate-free representation is utilized in engineering, in relativistic physics and in modeling the brain functions. In engineering, it was used to describe physical quantities in a way which is independent of the particular system of coordinates used for a given representation. In the context of engineering, parameter space is often three-dimensional and Euclidean and involves vector representations. Modeling the brain functions requires a generalization of tensor theory beyond low-dimensional Euclidean and curved spaces [49], [50]. In the present contribution, a further broadening of the types of spaces of interest in learning theory is invoked, because extended Hamiltonian learning theories do not necessarily relate to vector representations.

The present paper is organized as follows. Section II presents some general definitions and the notation used throughout the paper. It also presents the derivation of general equations of learning on Riemannian manifolds as well as the energy balance equation for a general extended Hamiltonian system. Section II also discusses the relationship between the presented learning theory and the classical learning theory, with particular reference to the 'momentum term' and the improvement of second-order learning with respect to first-order (i.e., gradient-based) learning. Section III discusses several cases of learning by extended Hamiltonian systems on manifolds of interest in the scientific literature (namely, the Stiefel manifold, the manifold of multidimensional rotations, the Grassmann manifold, the manifold of symmetric positive definite matrices, the generalized flag manifold and the manifold of real symplectic matrices). Section IV concludes the paper.

II. Extended Hamiltonian learning on Riemannian manifolds

In the present paper, both extrinsic and intrinsic coordinates will be used. In order to clarify the difference between extrinsic and intrinsic coordinates for manifold elements, consider the case of the space $\mathrm{SO}(2)$ of planar rotations. The manifold $\mathrm{SO}(2)$ is 1-dimensional, therefore, any element $x \in \mathrm{SO}(2)$ may be represented by one extrinsic coordinate: A matrix $x \in \mathrm{SO}(2)$ may be represented as

$$x = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \tag{3}$$

with $\theta \in [0\ 2\pi)$. On the other hand, by embedding the space $\mathrm{SO}(2)$ into the space $\mathbb{R}^{2\times 2}$, any element of $\mathrm{SO}(2)$ may be regarded as a 2-by-2 real-valued matrix

$$x = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}, \tag{4}$$

whose entries must satisfy the constrains:

$$\begin{cases} x_{11}^2 + x_{21}^2 = 1, \\ x_{22}^2 + x_{12}^2 = 1, \\ x_{11}x_{12} + x_{21}x_{22} = 0, \\ x_{11}x_{22} - x_{12}x_{21} = 1. \end{cases} \tag{5}$$

The four parameters x_{11}, x_{12}, x_{21}, x_{22} are termed *intrinsic* coordinates.

In the present section, extrinsic coordinates will be used in order to obtain a general expression for the learning dynamics on the smooth matrix-type manifold M and to derive the law of energy balance of the system. Although from a differential geometric perspective the choice of coordinates does not matter at all because all the expressions are covariant with respect to coordinate changes, from an implementation viewpoint the use of intrinsic coordinates and related expressions is more profitable. Therefore, in the section III, we shall abandon the extrinsic coordinates in favor of intrinsic ones and extended Hamiltonian systems on matrix-type manifolds of interest will be derived from the extended Hamilton principle in intrinsic coordinates, essentially by making use of the structure of the tangent spaces and normal spaces of embedded manifolds.

A. Definitions and notation

Let M be a differentiable manifold of dimension r. In a point $x \in M$, the tangent space to the manifold M is denoted as T_xM. Also, the notion of normal space of an embedded manifold $M \hookrightarrow \mathbb{R}^r$ may be defined as:

$$N_xM \stackrel{\mathrm{def}}{=} \{\zeta \in \mathbb{R}^r | \mathrm{tr}(\zeta^T v) = 0, \forall v \in T_xM\}, \tag{6}$$

where symbol $\mathrm{tr}(\cdot)$ denotes the trace of the matrix within.

Let the algebraic group (G, m, i, e) be a Lie group, namely, let G be endowed with a differentiable manifold structure, which is further supposed to be Riemannian. Here, operator $m : G \times G \to G$ denotes group multiplication, operator $i : G \to G$ denotes group inverse and $e \in G$ denotes group identity, namely $m(x, i(x)) = e$ for every $x \in G$. The algebraic and the differential-geometric structures need to be compatible, namely, the map $(x, y) \mapsto m(x, i(y))$ needs to be smooth for every $x, y \in G$. To the Lie group G, a Lie algebra $\mathfrak{g} \stackrel{\mathrm{def}}{=} T_eG$ is associated.

A Riemannian manifold M is endowed with a inner product $\langle\cdot,\cdot\rangle_x : T_xM \times T_xM \to \mathbb{R}$. The Euclidean metric is denoted by $\langle\cdot,\cdot\rangle^{\mathrm{E}}$. A local metric $\langle\cdot,\cdot\rangle_x$ also defines a local norm $\|v\|_x \stackrel{\mathrm{def}}{=} \sqrt{\langle v, v\rangle_x}$, for $v \in T_xM$.

Let $\psi : M \to \mathbb{R}$ denote a differentiable function. Symbol $\nabla_x\psi \in T_xM$ denotes the Riemannian gradient of function ψ with respect to a metric $\langle\cdot,\cdot\rangle_x$, which is defined by the compatibility condition

$$\langle \nabla_x\psi, v\rangle_x = \langle \partial_x\psi, v\rangle^{\mathrm{E}} \text{ for any } v \in T_xM. \tag{7}$$

In the following, an over-dot will denote the derivative $\frac{d}{dt}$, while a double over-dot will denote the derivative $\frac{d^2}{dt^2}$.

For the theory of differential manifolds and Lie groups, readers may refer to [60]. An introductory book on matrix manifolds is [1].

B. Hamilton and extended Hamilton principle: Law of motion and energy balance equation

A general principle to describe the dynamics of a conservative system on a differentiable manifold M is the Hamiltonian variational principle, that may be stated as:

$$\delta \int_{t_1}^{t_2} (K_x(\dot{x}, \dot{x}) - V) dt = 0. \tag{8}$$

In the above equation, $x(t) \in M$ denotes the trajectory of a massive particle that slides on the manifold M with an instantaneous velocity $\dot{x}(t) \in T_{x(t)}M$, the function $K_x : T_xM \times T_xM \to \mathbb{R}$ denotes the kinetic energy of the particle in a point $x \in M$ and the quantity $V : M \to \mathbb{R}$ denotes a potential energy field. All in one, the difference $K_x - V$ denotes the Lagrangian function associated to the particle, whose integral represents the total action associated to the particle. The time-interval $[t_1\ t_2] \subset \mathbb{R}$ denotes the time-span that the dynamics is observed within. The symbol δ denotes the change of the action associated to the system when moving from a point of the trajectory $x(t)$ to a point on an infinitely close trajectory corresponding to the

same value of the parameter t. For a reference on the calculus of variations see, e.g., [25].

The principle (8) establishes that the particle under observation moves over a trajectory that makes the total action stationary. Moreover, Hamilton's variational principle is suitable to describe the dynamics of conservative system, in that the solutions of the variational problem (8) make the total energy of the system preserve over time. The energy balance law for the system (8) reads:

$$K_x(\dot{x},\dot{x}) + V = \text{constant over the trajectory}. \tag{9}$$

In the ordinary space $\mathbb{R}^n$, a suitable generalization of the principle (8) to include non-conservative systems is stated as [58]:

$$\delta \int_{t_1}^{t_2} L(x,v)dt + \int_{t_1}^{t_2} f \cdot \delta x dt = 0, \tag{10}$$

where the function L denotes the Lagrangian function of the system, the vector v denotes instantaneous velocity, the vector f denotes a system of non-conservative forces and the symbol $\cdot$ denotes the ordinary inner product in $\mathbb{R}^n$. The extended Hamilton principle consists of the term in the left-hand side of equation (8) and a term that takes into account dissipation forces[1]. By calculating the variation of the leftmost integral and by recalling that the variation δx is arbitrary except that at the boundary of the trajectory where it vanishes, the extended version of the Euler-Lagrange equation is obtained:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial v}\right) - \frac{\partial L}{\partial x} = f. \tag{11}$$

The Euler-Lagrange equation for the system takes on the familiar form except for the term f on the right-hand side that takes into account the lack of conservation of energy due to dissipative forces. A conservative system has $f = 0$, which yields the familiar Euler-Lagrange equation. Moreover, according to [63], the principle (10) relates to the d'Alembert principle of virtual work.

On a Riemannian manifold M whose tangent spaces T_xM are endowed with the inner product $\langle\cdot,\cdot\rangle_x$, the extended Hamilton principle (10) may be further generalized to:

$$\int_{t_1}^{t_2} \delta(K_x(\dot{x},\dot{x}) - V)dt + \int_{t_1}^{t_2} \langle f_x, \delta x\rangle_x dt = 0, \tag{12}$$

where $f_x \in T_xM$ denotes a field of dissipation forces. On each point of the trajectory $t \in (t_1\ t_2)$, the variation $\delta x \in T_{x(t)}M$ is arbitrary, while at the boundaries of the trajectory it must vanish to zero. (Another possible generalization of the principle (8) would arise if the potential function is replaced by a generalized potential as in [35]. However, generalized potentials give rise only to a special instance of dissipation force field, namely, gyroscopic force fields.)

On a differential manifold M of dimension r, a point $x \in M$ may be described by extrinsic coordinates $(x^1,\dots,x^r)$ via a coordinate chart. Working with extrinsic coordinates eases the expression of the solution of the variational problem (12) in explicit form.

In order to expose the derivation of the general law of dynamics arising from the extended Hamilton principle (12), the following facts are worth summarizing:

- The standard notation for covariant and contravariant tensors indices as well as Einstein's convention on summation indices are made use of throughout the present section (unless otherwise stated).
- For $u, v \in T_xM$, it holds $\langle u, v\rangle_x = g_{ij}u^iv^j$, where g_{ij} denote the components of the metric tensor associated to the inner product $\langle\cdot,\cdot\rangle_x$. The functions g_{ij} depend on coordinates x^k and the symmetry property $g_{ij} = g_{ji}$ holds.
- In extrinsic coordinates, the Riemannian gradient of a regular function $f : M \to \mathbb{R}$ in a point $x \in M$ is given by:

$$(\nabla_x f)^h = g^{hk}\frac{\partial f}{\partial x^k}. \tag{13}$$

  The quantities $\frac{\partial f}{\partial x^k}$ denote the components of the Euclidean gradient, while functions g^{hk} denote the contravariant components of the covariant tensor field g_{hk}.
- The Christoffel symbols of the first kind associated to a metric tensor of components g_{ij} are defined as:

$$\Gamma_{kji} \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{\partial g_{ji}}{\partial x^k} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{kj}}{\partial x^i}\right).$$

  The Christoffel symbols of the second kind are defined as $\Gamma^h_{ik} \stackrel{\text{def}}{=} g^{hj}\Gamma_{ikj}$. The Christoffel symbols of the second kind are symmetric in the covariant indices, namely, $\Gamma^h_{ij} = \Gamma^h_{ji}$. The Christoffel form $\Gamma_x : T_xM \times T_xM \to \mathbb{R}^r$ is defined in extrinsic coordinates by $[\Gamma_x(v,w)]^k \stackrel{\text{def}}{=} \Gamma^k_{ij}v^iw^j$.
- The kinetic energy functional is the symmetric bilinear form $\frac{1}{2}\mathcal{M}\langle\dot{x},\dot{x}\rangle_x$, namely $K_x(\dot{x},\dot{x}) = \frac{1}{2}\mathcal{M}g_{ij}\dot{x}^i\dot{x}^j$, where $\mathcal{M} \geq 0$ plays the role of a mass term. On a Riemannian manifold, the metric tensor is positive-definite, hence $K_x(\dot{x},\dot{x}) \geq 0$ on every trajectory $x(t) \in M$. (Such property does not hold true on other manifolds of interest in neural learning, such as pseudo-Riemannian manifolds [22].)
- The potential energy function V depends on the coordinates x^k only.
- The variations δx^k may be chosen arbitrarily except that at the boundaries of the trajectory where they vanish, namely, $\delta x^k|_{t_1} = \delta x^k|_{t_2} = 0$.

In the present subsection as well as in subsection II-C, it is assumed that the dynamical system is massive, namely that $\mathcal{M} \neq 0$. The case of massless systems, namely, the case that $\mathcal{M} = 0$, is meaningful too in the present context and will be discussed in subsection II-D.

In extrinsic coordinates, the principle (12) may be rewritten as:

$$\int_{t_1}^{t_2} \delta\left(\frac{1}{2}\mathcal{M}g_{ij}\dot{x}^i\dot{x}^j - V\right)dt + \int_{t_1}^{t_2} f_k\delta x^k dt = 0, \tag{14}$$

where functions f_k denote the components of the covariant dissipation force field $g_{kh}f^h_x$. By calculating the variation in the leftmost integral, the extended Hamiltonian principle takes on the form:

$$\int_{t_1}^{t_2} \varphi_k \delta x^k dt = 0, \tag{15}$$

$$\varphi_k \stackrel{\text{def}}{=} \frac{1}{2}\mathcal{M}\frac{\partial g_{ij}}{\partial x^k}\dot{x}^i\dot{x}^j - \mathcal{M}\frac{d}{dt}(g_{ik}\dot{x}^i) + f_k - \frac{\partial V}{\partial x^k}, \tag{16}$$

which was calculated through integration by parts. As the variations δx^k may be chosen arbitrarily for $t \in (t_1\ t_2)$, the equations of dynamics on the manifold M are $\varphi_k = 0$, namely:

$$\frac{1}{2}\mathcal{M}\frac{\partial g_{ij}}{\partial x^k}\dot{x}^i\dot{x}^j - \mathcal{M}\frac{\partial g_{ik}}{\partial x^j}\dot{x}^i\dot{x}^j - \mathcal{M}g_{ik}\ddot{x}^i + f_k - \frac{\partial V}{\partial x^k} = 0. \tag{17}$$

By introducing the Christoffel symbols of the first kind Γ_{ijk}, the above equation may be rewritten as:

$$\mathcal{M}g_{ik}\ddot{x}^i + \mathcal{M}\Gamma_{ijk}\dot{x}^i\dot{x}^j + \frac{\partial V}{\partial x^k} - f_k = 0. \tag{18}$$

Multiplying both sides by the contravariant tensor g^{hk} and saturating with respect to the index k yields:

$$\mathcal{M}\ddot{x}^h + \mathcal{M}g^{hk}\Gamma_{ijk}\dot{x}^i\dot{x}^j + g^{hk}\left(\frac{\partial V}{\partial x^k} - f_k\right) = 0. \tag{19}$$

[1]The principle (10) may be used, e.g., to describe the dynamics of a viscous fluid where kinematic heat loss occurs. In a viscous fluid, kinematic heat loss gives rise to internal stress σ and the non-conservative force term f is proportional to $\frac{\partial\sigma}{\partial x}$ [58].

In conclusion, the extended Hamiltonian equation (19) may be rewritten as the system:

$$\left\{ \begin{array}{lcl} \dot{x} & = & \frac{1}{\mathcal{M}} v, \\ \dot{v} & = & -\mathcal{M}\Gamma_x(v,v) - \nabla_x V + f_x. \end{array} \right. \tag{20}$$

In the present manuscript, the dissipation term is drawn from viscosity theory, namely, $f_x \stackrel{\text{def}}{=} -\mu\dot{x}$, with $\mu \geq 0$. The learning dynamics assumes then the final expression:

$$\left\{ \begin{array}{lcl} \dot{x} & = & \frac{1}{\mathcal{M}} v, \\ \dot{v} & = & -\mathcal{M}\Gamma_x(v,v) - \nabla_x V - \mu v, \end{array} \right. \tag{21}$$

The dynamics (21) is governed by a driving force from a potential energy plus a viscosity term. The Christoffel part arises from the curveness of the manifold-type parameter space. In particular, the second equation may be interpreted as a law of evolution of the velocity vector v along the trajectory followed by the extended Hamiltonian system.

The law of energy balance for the dissipative system (21) may be found by the equation:

$$\int_{t_1}^{t} \varphi_k \dot{x}^k dt = 0, \tag{22}$$

which holds for every $t \in [t_1 \ t_2]$. Plugging in the expression (16) gives:

$$\begin{array}{c} \frac{1}{2}\mathcal{M}\int_{t_1}^{t} \frac{\partial g_{ij}}{\partial x^k}\dot{x}^k\dot{x}^i\dot{x}^j dt - \mathcal{M}\int_{t_1}^{t} \frac{d}{dt}(g_{ik}\dot{x}^i)\dot{x}^k dt \\ -\mu\int_{t_1}^{t} \dot{x}_k\dot{x}^k dt - \int_{t_1}^{t} \frac{\partial V}{\partial x^k}\dot{x}^k dt = 0. \end{array} \tag{23}$$

Adopting the short notations $K(t) \stackrel{\text{def}}{=} K_{x(t)}(\dot{x}(t), \dot{x}(t))$ and $V(t) \stackrel{\text{def}}{=} V(x(t))$, the energy balance equation (23) may be written as:

$$K(t) + V(t) = V(t_1) + K(t_1) - 2\frac{\mu}{\mathcal{M}}\int_{t_1}^{t} K(t)dt. \tag{24}$$

It shows that the system looses energy at a rate proportional to its kinetic energy and to the viscosity index μ.

C. Free motion on geodesics and inertial learning

On a connected Riemannian manifold, it is customary to define special curves that represent the counterparts of straight paths over flat manifolds. Any of these special curves is termed geodesic arc and is defined as the path of minimal length connecting two assigned points on a manifold.

In terms of the calculus of variations, a geodesic on a manifold M is defined as the curve $x(t) \in M$, $t \in [t_1 \ t_2]$ such that:

$$\int_{t_1}^{t_2} \delta\langle \dot{x}, \dot{x}\rangle_x^{\frac{1}{2}} dt = 0. \tag{25}$$

By recalling that the kinetic energy for a particle is defined as $K_x(\dot{x}, \dot{x}) = \frac{1}{2}\mathcal{M}\langle \dot{x}, \dot{x}\rangle_x$, the variational principle (25) takes on the form:

$$\int_{t_1}^{t_2} \delta\sqrt{2K_x} dt = 0, \tag{26}$$

with $2K_x = \mathcal{M} g_{ij}\dot{x}^i\dot{x}^j$. Computing the variation within the integral yields the equation:

$$\int_{t_1}^{t_2} \frac{1}{\sqrt{K_x}} \left(\frac{\partial g_{ij}}{\partial x^k}\dot{x}^i\dot{x}^j + 2g_{ik}\dot{x}^i\frac{d}{dt}\delta x^k \right) \delta x^k dt = 0.$$

Integration by part yields the variational equation for the geodesic:

$$\int_{t_1}^{t_2} \varphi_k \delta x^k dt = 0, \tag{27}$$

$$\varphi_k = \frac{1}{2\sqrt{K_x}}\frac{\partial g_{ij}}{\partial x^k}\dot{x}^i\dot{x}^j - \frac{d}{dt}\left(\frac{g_{ik}\dot{x}^i}{\sqrt{K_x}}\right). \tag{28}$$

As the variations δx^k may be chosen arbitrarily, the equation of the geodesic must be $\varphi_k = 0$.

There exists a parametrization, termed 'normal parametrization', that keeps the quantity $\sqrt{K_x}$, namely $\|\dot{x}\|_x$, constant over the trajectory $x(t)$. Under normal parametrization, the equation of the geodesic simplifies into:

$$\frac{d}{dt}\left(g_{ik}\dot{x}^i\right) - \frac{1}{2}\frac{\partial g_{ij}}{\partial x^k}\dot{x}^i\dot{x}^j = 0. \tag{29}$$

By reasoning as in the subsection II-B, the above equation may be expanded and written in compact form as:

$$\ddot{x} + \Gamma_x(\dot{x}, \dot{x}) = 0. \tag{30}$$

The above second-order differential equation needs two initial conditions to be solved. Whenever the initial conditions are $x(0) = x \in M$ and $\dot{x}(0) = v \in T_x M$, the solution of the geodesic equation will be denoted by $c_{x,v}(t) \in M$.

Comparing the expression (30) with the extended Hamiltonian system (21), it is immediate to see that they coincide for a non-dissipative system ($\mu = 0$) which is clueless to learning (namely, whose goal function V does not vary over the manifold M). In fact, in this case the variational learning principle (12) coincides with the variational principle (26); moreover, the Hamiltonian formulation produces a trajectory that keeps the quantity that undergoes variation, namely the only kinetic energy in this case, constant over the trajectory itself. The Christoffel form $\Gamma_x(\cdot,\cdot)$ may thus be computed on the basis of the variational principle:

$$\int_{t_1}^{t_2} \delta K_x dt = 0. \tag{31}$$

It is worth noting that, on a Riemannian manifold, the Christoffel matrix-function $\Gamma_x(\cdot,\cdot)$ depends only on the metric tensor, therefore, the variational principle (31) may be used to compute the Christoffel matrix regardless of the type of motion at hand (i.e., free motion on a geodesic curve or forced motion on a non-geodesic arc).

A "clueless" learning system proceeds under the only constraints imposed by the manifold structure and by the chosen metric along the shortest-length learning curve over the manifold. On a flat parameter space, this would correspond to a system that learns though a straight path with constant speed, which we might refer to as "inertial learning".

It is customary to assume the interval $[t_1 \ t_2]$ for the normal parametrization of a geodesic arc as $[0 \ 1]$.

As a note on method, the equation of geodesics as well as the Hamiltonian learning equations might be derived in terms of covariant derivative instead of calculus of variations on manifold. We chose the latter type of derivation because it proves more convenient when calculating Christoffel operator and geodesic forms in intrinsic coordinates.

D. On the relationships with classical learning theory

The quantities that appear in the extended Hamiltonian system of equations (21) as well as in the energy balance equation (24) may be given a meaning in light of the formulation of a general theory of learning on Riemannian manifolds.

The present subsection is devoted to an analysis of the relationships between the discussed learning theory and the classical learning theory. In particular, two aspects will be discussed in details:

- A relationship between dynamical learning and gradient-based steepest-descent learning as well as the emergence of resemblance between dynamical learning and gradient learning with a 'momentum' term.

- A convergence analysis of dynamical learning on manifolds in terms of Lyapunov functions as well as in terms of behavior analysis in the proximity of an optimal pattern.

In order to expose the above topics, it is worth recalling the differential-geometric notion of *normal coordinate system*, which plays an important role in the subsequent analysis.

In a given point $x \in M$ of a differentiable manifold M equipped with a symmetric affine connection, normal coordinates are defined as an extrinsic coordinate system in an open neighborhood of the point x obtained by applying a specific map to the tangent space T_xM. It is worth recalling the following noticeable properties:

- In a normal coordinate system of an open neighborhood of $x \in M$, the Christoffel symbols vanish at the point x, namely, $\Gamma_{ij}^k\big|_x = 0$.
- It is possible to choose a normal coordinate system associated to a Riemannian manifold so that the metric tensor equals the identity tensor at the point x, namely, $g_{ij}\big|_x = 0$ for $i \neq j$ and $g_{ii}\big|_x = 1$.

Both properties help simplifying calculations in extrinsic coordinates. It is worth noting that if a normal coordinate system is chosen in an open neighborhood of a point x such that the metric tensor g_{ij} equals the identity tensor at the point x, the same holds for the dual metric tensor g^{ij}.

In the present subsection, it is assumed that the learning system is dissipative, namely, that $\mu \neq 0$.

1) Comparison of extended Hamiltonian learning and gradient steepest descent: When a manifold of interest M and a differentiable learning goal function $V : M \to \mathbb{R}$ are specified, a known way to set up a learning rule refers to the Riemannian gradient steepest descent optimization algorithm, that may be expressed by the system:

$$\dot{x} = -\epsilon \nabla_x V, \tag{32}$$

where the parameter $\epsilon > 0$ is termed 'learning stepsize'. The flow $x(t)$ associated to system (32) tends toward a local minimum of the potential energy V, in fact:

$$\dot{V} = \langle \nabla_x V, \dot{x} \rangle_x = -\epsilon \|\nabla_x V\|_x^2 \leq 0, \tag{33}$$

with equality holding if and only if $\nabla_x V = 0$, namely, when the flow $x(t)$ approaches a stationary point of the learning goal V.

Now, the dynamics (21) on the manifold M takes on the form:

$$\mathcal{M}\ddot{x} + \mu\dot{x} + \mathcal{M}\Gamma_x(\dot{x}, \dot{x}) = -\nabla_x V. \tag{34}$$

A massless dynamical system is characterized by the condition $\mathcal{M} = 0$. Setting to zero the mass parameter in the dynamical equation (34) yields:

$$\dot{x} = -\frac{1}{\mu}\nabla_x V. \tag{35}$$

Hence, the extended Hamiltonian learning equation (34) for a massless dynamical system collapses to the Riemannian-gradient-steepest-descent learning equation (32) with learning stepsize $\epsilon = \frac{1}{\mu}$.

For a massive dynamical system, some differences between the extended Hamiltonian system (34) and the gradient steepest descent rule (32) are immediately recognizable. In the case of the extended Hamiltonian learning, for the same initial point $x(0)$, different choices of the initial velocity $\dot{x}(0)$ may provide additional control of the expected solution (namely, second-order learning provides more degrees of freedom with respect to gradient-based one [2], [10]).

The term $\mathcal{M}\Gamma_x(\dot{x}, \dot{x}) + \mu\dot{x}$ contains information on learning speed and it is reminiscent of the 'momentum term' in classical learning theory.

Let us recall the notion of momentum term in classical learning theory. Numerically, on an parameter space $\mathbb{R}^n$, an integration scheme for the learning equation (32) is:

$$x(t + \Delta t) - x(t) = -\epsilon \left.\frac{\partial V}{\partial x}\right|_{x(t)} + \beta[x(t) - x(t - \Delta t)], \tag{36}$$

where, by a slight abuse of notation, the same symbol to denote the continuous-time flow $x(t)$ was used to denote its sampled version, symbol Δt denotes a sampling interval and $\beta \in \mathbb{R}$ is termed 'momentum parameter'. It is convenient to introduce the parameter-change

$$\Delta x^k(t) \stackrel{\text{def}}{=} x^k(t + \Delta t) - x^k(t), \tag{37}$$

for the kth component x^k of the parameter vector x. The equation (36) becomes, then:

$$\Delta x^k(t) = -\epsilon \left.\frac{\partial V}{\partial x^k}\right|_{x^k(t)} + \beta \Delta x^k(t - \Delta t). \tag{38}$$

The 'momentum term' $\beta\Delta x^k(t - \Delta t)$ was included to take into account the previous change in the parameter and was numerically found to be able to improve the convergence speed (for a reference and historical notes, see [52]).

By making use of extrinsic coordinates, the dynamical system (34) on a r-dimensional Riemannian manifold M may be shown to resemble the classical equation (38). In extrinsic coordinates, the learning system (34) may be written as:

$$\mathcal{M}\ddot{x}^k + \mu\dot{x}^k + \mathcal{M}\Gamma_{ij}^k \dot{x}^i \dot{x}^j + (\nabla_x V)^k = 0. \tag{39}$$

The above differential equation may be discretized on $\mathbb{R}^r$ by invoking the approximations:

$$\dot{x}^k(t) \approx \frac{x^k(t + \Delta t) - x^k(t)}{\Delta t}, \tag{40}$$

$$\ddot{x}^k(t) \approx \frac{x^k(t + \Delta t) - 2x^k(t) + x^k(t - \Delta t)}{(\Delta t)^2}. \tag{41}$$

By plugging the above approximations in the equation (39), the left-hand side of the equation (39) becomes:

$$\begin{aligned} &\tfrac{\mathcal{M}}{(\Delta t)^2}[\Delta x^k(t) - \Delta x^k(t - \Delta t)] + \tfrac{\mu}{\Delta t}\Delta x^k(t) \\ &\quad + \tfrac{\mathcal{M}}{(\Delta t)^2}\Gamma_{ij}^k \Delta x^i(t) \Delta x^j(t) + (\nabla_x V)^k. \end{aligned} \tag{42}$$

By choosing a normal coordinate system in an open neighborhood of the point $x(t)$, the Christoffel symbols vanish at $x(t)$ and the approximate discrete-time version of equation (39) simplifies into:

$$\Delta x^k(t) = -\frac{(\Delta t)^2}{\mathcal{M} + \mu\Delta t}(\nabla_x V)^k + \frac{\mathcal{M}}{\mathcal{M} + \mu\Delta t}\Delta x^k(t - \Delta t). \tag{43}$$

The expression (43) resembles the expression (38) upon the following identifications:

$$\text{Momentum parameter } \beta \text{ with } \frac{\mathcal{M}}{\mathcal{M} + \mu\Delta t}, \tag{44}$$

$$\text{Learning stepsize } \epsilon \text{ with } \frac{(\Delta t)^2}{\mathcal{M} + \mu\Delta t}. \tag{45}$$

Informally, the rationale for the inclusion of the momentum term in Euclidean-gradient-steepest-descent is that gradient-steepest-descent learning slows down when the trajectory $x(t)$ enters a long narrow valley in the surface of the criterion function V. In this case, the direction of the gradient $\frac{\partial V}{\partial x}$ is almost perpendicular to the long axis of the valley and thus the trajectory oscillates perpendicularly to the long axis of the valley and only slightly moves out of the valley itself. The momentum term has the effect of averaging out the oscillations and to add up a contribution along the long axis, hence helping to mitigate the plateau effect.

2) Lyapunov function and fine-convergence analysis: Recall that the quantity x denotes the set of adaptable parameters in a neural system. The function $x(t)$ denotes a learning trajectory over the manifold M. The function $\dot{x}(t)$ denotes the instantaneous learning speed and direction. Likewise, the function $\ddot{x}(t)$ denotes the instantaneous learning acceleration. The function V denotes a learning goal. If the regular function $V : M \to \mathbb{R}$ possesses a minimum $V_{\rm m}$ in M, define the function:

$$Y(t) \stackrel{\rm def}{=} K(t) + V(t) - V_{\rm m}. \tag{46}$$

As the kinetic energy is a positive-definite form, namely $K(t) \geq 0$, with equality if and only if $\dot{x} = 0$, it follows that $Y(t) \geq 0$. Moreover, from equation (24) it follows that:

$$\dot{Y}(t) = -\frac{2\mu}{\mathcal{M}} K(t) \leq 0, \tag{47}$$

with equality if and only if $\dot{x} = 0$. Hence the function $Y(t)$ is a Lyapunov function for the system (21). The function $Y(t)$ thus tends asymptotically to 0, which implies that the function V tends asymptotically to the (locally) minimum value $V_{\rm m}$. In other words, the function V represents a cost-type learning goal as the neural system learns to minimize the potential energy.

The term $-\mu\dot{x}$ in the extended Hamiltonian system makes the dynamics be subjected to viscous drag (see for instance [2], [14]), stabilizes the learning dynamics and may prevent fluctuations for sufficiently high viscosity values.

In the case of dynamical learning over the manifold $M = \mathbb{R}^n$, an analysis of the convergence properties and of the improvement due to dynamical learning over gradient-based learning was conducted in [52]. Such analysis is based on the local quadratic approximation of the learning criterion function around a local minimum. We now aim at extending such study to the case of a curved Riemannian manifold.

In extrinsic coordinates, the learning system (34) may be rewritten as:

$$\mathcal{M}\ddot{x}^k + \mu\dot{x}^k + \mathcal{M}\Gamma^k_{ij}\dot{x}^i\dot{x}^j + g^{ki}\frac{\partial V}{\partial x^i} = 0. \tag{48}$$

Let $(x^1_{\rm m}, \ldots, x^r_{\rm m})$ denote the coordinates of a local minimum $x_{\rm m} \in M$ of the potential energy function V. The potential may be approximated by Taylor expansion on an open neighborhood of the above local minimum as:

$$V = V_{\rm m} + \left.\frac{\partial V}{\partial x^k}\right|_{\rm m} (x^k - x^k_{\rm m}) + \frac{1}{2}\left.\frac{\partial^2 V}{\partial x^i \partial x^j}\right|_{\rm m} (x^i - x^i_{\rm m})(x^j - x^j_{\rm m}) + \cdots, \tag{49}$$

where symbol $|_{\rm m}$ denotes evaluation in the point $(x^1_{\rm m}, \ldots, x^r_{\rm m})$. In a local minimum, it holds $\left.\frac{\partial V}{\partial x^k}\right|_{\rm m} = 0$ and the coefficients $h_{ij}|_{\rm m} \stackrel{\rm def}{=} \left.\frac{\partial^2 V}{\partial x^i \partial x^j}\right|_{\rm m}$ arrange in a Hessian matrix H which is symmetric and positive-definite. Therefore, a first-order approximation of the differential equation (48) is:

$$\begin{aligned} &\mathcal{M}\tfrac{d^2}{dt^2}(x^k - x^k_{\rm m}) + \mu\tfrac{d}{dt}(x^k - x^k_{\rm m}) \\ &+\mathcal{M}\,\Gamma^k_{ij}\big|_{\rm m}\,\tfrac{d}{dt}(x^i - x^i_{\rm m})\tfrac{d}{dt}(x^j - x^j_{\rm m}) \\ &+ g^{ki}\big|_{\rm m}\, h_{ij}|_{\rm m}\,(x^j - x^j_{\rm m}) = 0. \end{aligned} \tag{50}$$

By choosing a normal coordinate system in an open neighborhood of the point $x_{\rm m} \in M$, the above equation simplifies noticeably as all coefficients $\Gamma^k_{ij}\big|_{\rm m}$ vanish to zero. In addition, the normal coordinate system may be chosen in a way that $g^{ki}\big|_{\rm m}$ equals the identity. Therefore, the differential equation (50) simplifies into:

$$\mathcal{M}\frac{d^2}{dt^2}(x^k - x^k_{\rm m}) + \mu\frac{d}{dt}(x^k - x^k_{\rm m}) + h_{ki}|_{\rm m}\,(x^i - x^i_{\rm m}) = 0. \tag{51}$$

Now, the system of differential equations (51) decouples via diagonalization of the matrix H. As the matrix H is symmetric and positive-definite, it decomposes as $H = Q\Lambda Q^T$, where Q denotes an orthogonal $r \times r$ matrix and $\Lambda \stackrel{\rm def}{=} {\rm diag}(\lambda_1, \ldots, \lambda_r)$ with all $\lambda_k > 0$. The variable change:

$$(z^1, \ldots, z^r) \stackrel{\rm def}{=} Q(x^1 - x^1_{\rm m}, \ldots, x^r - x^r_{\rm m}) \tag{52}$$

decouples the differential system (51) into:

$$\mathcal{M}\ddot{z}^k + \mu\dot{z}^k + \lambda_k z^k = 0, \tag{53}$$

where it is understood that the Einstein summation convention does not hold.

The differential system (53) is formally equivalent to the one studied in [52], hence the analysis of the improvement due to second-order dynamics over first-order dynamics in approaching a local minimum of the potential energy function leads to the same conclusions of paper [52]. Such conclusions may be summarized as follows.

For a massless system, that corresponds to gradient-steepest descent learning, it holds $\mathcal{M} = 0$, hence the convergence of the variable z^k to zero is of the type:

$$\exp\left(-\frac{\lambda_k}{\mu}t\right). \tag{54}$$

For a massive system (namely, $\mathcal{M} \neq 0$), each differential equation (53) represents a (damped) harmonic oscillator. In the case of a massive system, the convergence of the variable z^k to zero is thus dominated by the term:

$$\exp\left(-\left|{\rm Re}\left\{-\frac{\mu}{2\mathcal{M}} + \sqrt{\frac{\mu}{\mathcal{M}}\left(\frac{\mu}{4\mathcal{M}} - \frac{\lambda_k}{\mu}\right)}\right\}\right| t\right), \tag{55}$$

where symbol ${\rm Re}$ denotes real part.

Comparing the speed of convergence of the above two terms, one reaches the conclusion that for those variables z^k for which the condition

$$\lambda_k < \frac{\mu^2}{2\mathcal{M}} \tag{56}$$

holds true, the dynamical system converges faster than the gradient-steepest descent one.

Manifolds of interest in the scientific literature may be compact as well as non-compact. Regular functions on compact manifolds possess a minimum while regular functions whose variables take values on non-compact manifolds do not always have a minimum even if they are bounded from below. The soundness of a given optimization problem is a pre-requisite to the application of an optimization algorithm.

III. Dynamics on special manifolds

The present section discusses in details some cases of interest, namely, dynamics on the Stiefel manifold equipped with the Euclidean metrics and on the Stiefel manifold equipped with the canonical metrics, dynamics over the special orthogonal group, dynamics over the Grassmann manifold, dynamics over the group of symmetric positive definite matrices, dynamics over the generalized flag manifold and dynamics over the real symplectic group of matrices.

Throughout the present section, the mass parameter will be considered unitary, namely, $\mathcal{M} = 1$, for the ease of notation.

A. Dynamics on the compact Stiefel manifold

The (compact) Stiefel manifold is defined as:

$${\rm St}(n,p) = \{x \in \mathbb{R}^{n\times p} | x^T x = e_p\}, \tag{57}$$

where $p \leq n$, superscript ^{T} denotes matrix transpose and symbol e_p denotes a $p \times p$ identity matrix. Given a trajectory $x(t) \in {\rm St}(n,p)$, derivation with respect to the time parameter yields $\dot{x}^T x + x^T\dot{x} = 0$,

which means that the tangent space to the manifold $\mathrm{St}(n,p)$ in a point $x \in \mathrm{St}(n,p)$ has structure:

$$T_x\mathrm{St}(n,p) = \{v \in \mathbb{R}^{n\times p} | v^T x + x^T v = 0\}. \tag{58}$$

The normal space has structure:

$$N_x\mathrm{St}(n,p) = \{xs | x \in \mathbb{R}^{n\times p},\ s^T - s = 0\}. \tag{59}$$

Exemplary learning problems where the orthogonality constraint plays a prominent role are signal representation by principal/minor component analysis on the Stiefel manifold [48], [65], blind source separation upon signal pre-whitening and independent component analysis [7], [33], [45], non-negative matrix factorization [66], direction of arrival estimation [40], best basis search/selection [4], [34], [44], electronic structures computation within local density approximation, e.g. for understanding the thermodynamics of bulk materials, the structure and dynamics of surfaces, and the nature of point-defects in crystals [12] and factor analysis in psychometrics [13]. Moreover, the compact singular value decomposition theorem provides a widely-known mathematical tool that allows one to recast any linear problem into a pair of Stiefel learning problems.

1) Dynamics on the Stiefel manifold with Euclidean metric: A possible metric that the Stiefel manifold may be endowed with is the Euclidean metric:

$$\langle u, v\rangle_x = \mathrm{tr}(u^T v),\ u, v \in T_x\mathrm{St}(n,p). \tag{60}$$

As seen in the subsection II-C, the symmetric form $\Gamma_x(\dot{x},\dot{x})$ may be computed by the variational principle (31) customized to the chosen metric, namely:

$$\int_0^1 \delta \mathrm{tr}(\dot{x}^T\dot{x})dt = 0. \tag{61}$$

Straightforward calculations show that:

$$\int_0^1 \delta \mathrm{tr}(\dot{x}^T\dot{x})dt = 2\int_0^1 \mathrm{tr}\left(\dot{x}^T \frac{d\delta x}{dt}\right)dt = -2\int_0^1 \mathrm{tr}(\ddot{x}^T\delta x)dt,$$

where the last equality holds upon integration by parts and where the variation δx corresponds to moving from a point on the curve $x(t)$ to a point corresponding to the same time t on a infinitely close curve on $\mathrm{St}(n,p)$; as a consequence, the variation $\delta x \in T_x\mathrm{St}(n,p)$. The equation of the geodesic in intrinsic form is thus such that, for every arbitrary $\delta x \in T_x\mathrm{St}(n,p)$ it holds $\mathrm{tr}(\ddot{x}^T\delta x) = 0$. This condition is verified if $\ddot{x} \in N_x\mathrm{St}(n,p)$, namely, if $\ddot{x} = xs$, with $s \in \mathbb{R}^{p\times p}$ being such that $s^T = s$. By deriving twice the condition $x^T x = e$, it follows $\ddot{x}^T x + x^T\ddot{x} + 2\dot{x}^T\dot{x} = 0$. Plugging the equation $\ddot{x} = -xs$ into the above equation yields $s = -\dot{x}^T\dot{x}$, so that the geodesic equation becomes $\ddot{x} + x\dot{x}^T\dot{x} = 0$, hence:

$$\Gamma_x(v,v) = x(v^T v),\ v \in T_x\mathrm{St}(n,p). \tag{62}$$

Note that the Christoffel form $\Gamma_x(\cdot,\cdot)$ was derived in intrinsic coordinates directly, without any need to resort to extrinsic coordinates-components. The solution of the geodesic equation (30) with the Christoffel matrix-function (62), with initial conditions $x(0) = x \in \mathrm{St}(n,p)$ and $\dot{x}(0) = v \in T_x\mathrm{St}(n,p)$, reads [12]:

$$c_{x,v}(t) = [x\ v]\exp\left(t\begin{bmatrix} x^T v & -v^T v \\ e_p & x^T v \end{bmatrix}\right) e_{2p,p}\exp(-tx^T v), \tag{63}$$

where symbol $\exp(\cdot)$ denotes matrix exponential. Also, it is interesting to verify that the matrix $\Gamma_x(v,v)$ belongs to the normal space $N_x\mathrm{St}(n,p)$ at any point $x \in \mathrm{St}(n,p)$ and for every tangent direction $v \in T_x\mathrm{St}(n,p)$, in fact, it holds $\mathrm{tr}(v^T\Gamma_x(v,v)) = 0$.

In order to complete the extended Hamiltonian system (21), it is necessary to compute the Riemannian gradient $\nabla_x V$. The Riemannian gradient of a function $V : \mathrm{St}(n,p) \to \mathbb{R}$ at a point $x \in \mathrm{St}(n,p)$ is the unique matrix in $T_x\mathrm{St}(n,p)$ such that:

$$\mathrm{tr}\left(u^T\partial_x V\right) = \langle u, \nabla_x V\rangle_x, \forall u \in T_x\mathrm{St}(n,p). \tag{64}$$

The above condition becomes $\mathrm{tr}\left(u^T\left(\partial_x V - \nabla_x V\right)\right) = 0$, which implies $\nabla_x V = \partial_x V + xs$, with s symmetric. Pre-multiplying this equation by matrix x^T yields $s + x^T\partial_x V = x^T\nabla_x V$. Transposing both hands of the above equation and summing hand-by-hand yields:

$$s = -\frac{1}{2}\left(\partial_x^T V x + x^T\partial_x V\right) + \frac{1}{2}\left(\nabla_x^T V x + x^T\nabla_x V\right). \tag{65}$$

As $\nabla_x V \in T_x\mathrm{St}(n,p)$, according to equation (58), it holds $\nabla_x^T V x + x^T\nabla_x V = 0$. In conclusion, the sought-for Riemannian gradient reads:

$$\nabla_x V = \partial_x V - \frac{1}{2}x\left(\partial_x^T V x + x^T\partial_x V\right). \tag{66}$$

2) Dynamics on the Stiefel manifold with canonical metric: The Stiefel manifold may be endowed with a second kind of metric, termed 'canonical metric'. The associated inner product reads:

$$\langle u, v\rangle_x = \mathrm{tr}\left(u^T\left(e_n - \frac{1}{2}xx^T\right)v\right),\ u, v \in T_x\mathrm{St}(n,p), \tag{67}$$

which, unlike the Euclidean metric (60), is not uniform over the Stiefel manifold.

Inserting the expression of the new kinetic energy form within the variational principle (31), computing the variation and integrating by parts as already done in the subsection III-A1, yields:

$$\frac{d}{dt}\left(\left(e_n - \frac{1}{2}xx^T\right)\dot{x}\right) + \frac{1}{2}\dot{x}\dot{x}^T x = xs,$$
$$s = -\dot{x}^T\dot{x} - (x^T\dot{x})^2.$$

Expanding the above expressions yields the following Christoffel function:

$$\Gamma_x(v,v) = -vv^T x - xv^T(e_n - xx^T)v,\ v \in T_x\mathrm{St}(n,p). \tag{68}$$

According to [12], the geodesic arc which is the solution of the geodesic equation (30), with the Christoffel matrix-function (68) with boundary conditions $x(0) = x \in \mathrm{St}(n,p)$ and $\dot{x}(0) = v \in T_x\mathrm{St}(n,p)$, may be computed as follows. Let q and r denote the factors of the compact QR decomposition of the matrix v, then:

$$c_{x,v}(t) = [x\ q]\exp\left(t\begin{bmatrix} x^T v & -r^T \\ r & 0_p \end{bmatrix}\right)\begin{bmatrix} e_p \\ 0_p \end{bmatrix}. \tag{69}$$

Again, it is interesting to verify that the matrix $\Gamma_x(v,v)$ given by equation (68) belongs to the normal space $N_x\mathrm{St}(n,p)$ at any point $x \in \mathrm{St}(n,p)$ and for every tangent direction $v \in T_x\mathrm{St}(n,p)$, in fact:

$$\mathrm{tr}(v^T\Gamma_x(v,v)) =$$
$$-\mathrm{tr}(v^T xv^T v) - \mathrm{tr}(v^T vv^T x) - \mathrm{tr}(v^T x(x^T v)^2) = 0.$$

The Riemannian gradient of a function $V : \mathrm{St}(n,p) \to \mathbb{R}$ at a point $x \in \mathrm{St}(n,p)$ with the metric (68) is the unique matrix $\nabla_x V$ in $T_x\mathrm{St}(n,p)$ such that:

$$\mathrm{tr}\left(u^T\left(\partial_x V - \left(e_n - \tfrac{1}{2}xx^T\right)\nabla_x V\right)\right) = 0,$$
$$\forall u \in T_x\mathrm{St}(n,p),$$

namely:

$$\partial_x V - \left(e_n - \frac{1}{2}xx^T\right)\nabla_x V = xs,$$
$$s = \frac{1}{2}\left(x^T\partial_x V + \partial_x^T V x\right).$$

Solving for the Riemannian gradient yields the final expression:

$$\nabla_x V = \partial_x V - x\partial_x^T V x. \tag{70}$$

Straightforward calculations show that the kinetic energy assumes the expression:

$$K_x(v,v) = \frac{1}{2}\mathrm{tr}(v^T v) + \frac{1}{4}\mathrm{tr}((v^T x)^2),\ v \in T_x\mathrm{St}(n,p). \tag{71}$$

B. Dynamics on the unit hypersphere

The unit hypersphere $\mathrm{S}^{n-1} = \left\{x \in \mathbb{R}^n | x^T x = 1\right\}$ coincides with the Stiefel manifold $\mathrm{St}(n,p)$ with $p = 1$. Moreover, on the Stiefel manifold $\mathrm{St}(n,1)$, the Euclidean metric and the canonical metric coincide, namely:

$$\langle u, v\rangle_x = u^T v,\ \forall u, v \in T_x\mathrm{S}^{n-1}, \tag{72}$$

where:

$$T_x\mathrm{S}^{n-1} = \{v \in \mathbb{R}^n | v^T x = 0\}. \tag{73}$$

There is a number of neural signal processing algorithms that learn parameter-vectors on the manifold S^{n-1} as, for instance, blind deconvolution algorithms [17], [20], [57], robust constrained beamforming algorithms [16], algorithms for data classification by linear discrimination based on non-gaussianity discovery [47] and algorithms for the maximization of the weighted sum rate in the MIMO broadcast channel under linear filtering [32].

The norm associated to the inner product (72) is the standard vector norm $\|\cdot\|$ and the Christoffel symmetric form may be customized as:

$$\Gamma_x(v,v) = -\|v\|^2 x,\ v \in T_x\mathrm{S}^{n-1}, \tag{74}$$

which clearly represents a radial term, hence normal to the surface of the hypersphere at any point. The associated geodesic may be given a closed-form expression, for $v \neq 0$, that is:

$$c_{x,v}(t) = x\cos(\|v\|t) + v\|v\|^{-1}\sin(\|v\|t), \tag{75}$$

which represents a great circle on the hypersphere, while $c_{x,0}(t) = x$.

The Riemannian gradient $\nabla_x V$ has the same expression as in equation (70), that may be also rewritten, in the case of the manifold S^{n-1}, as:

$$\nabla_x V = (e_n - xx^T)\partial_x V. \tag{76}$$

Such expression may be given the following interpretation: The Riemannian gradient on the real hypersphere S^{n-1} computes as the orthogonal projection of the Euclidean gradient $\partial_x V$ onto the tangent hyperplane $T_x\mathrm{S}^{n-1}$ by means of the projector $\mathcal{P}_x \stackrel{\mathrm{def}}{=} e_n - xx^T$.

C. Dynamics on the special orthogonal group

A number of neural-network applications and machine-learning applications deal with special-orthogonal-group connection patterns and their merging, like, for instance invariant visual perception [54], modeling of DNA chains [31], [41], automatic object pose estimation [61], kernel machines [64], study of plate tectonics [51], as well as blind source separation and independent component analysis [38]. For a recent review of other applications, readers might want to see, e.g., [18].

The special orthogonal group:

$$\mathrm{SO}(n) = \left\{x \in \mathbb{R}^{n\times n} | x^T x = e_n,\ \det(x) = 1\right\}, \tag{77}$$

coincides with the Stiefel manifold $\mathrm{St}(n,p)$ with $p = n$ with the extra constraint that its elements have positive determinant and may be thought of as a set of continuous high-dimensional rotations. Moreover, the manifold $\mathrm{SO}(n)$ is a Lie group with Lie algebra:

$$\mathfrak{so}(n) = \{\omega \in \mathbb{R}^{n\times n} | \omega^T + \omega = 0\}. \tag{78}$$

On Lie groups, solving the extended Hamiltonian learning system is simpler, essentially because all tangent spaces to the manifold may be described in terms of translation of the tangent space at identity, namely, the Lie algebra. In the case of the Lie group $\mathrm{SO}(n)$, the following identity holds:

$$T_x\mathrm{SO}(n) = \{x\omega | \omega \in \mathfrak{so}(n)\}. \tag{79}$$

On the special orthogonal group, the Stiefel manifold's Euclidean metric and canonical metric coincide up to an inessential constant factor $\frac{1}{2}$, therefore it is customary to set:

$$\langle x\omega_u, x\omega_v\rangle_x = -\mathrm{tr}(\omega_u\omega_v), \tag{80}$$

with $\omega_u, \omega_v \in \mathfrak{so}(n)$. The Christoffel symmetric form writes then:

$$\Gamma_x(x\omega, x\omega) = -x\omega^2,\ \omega \in \mathfrak{so}(n), \tag{81}$$

while the geodesic curve associated to the above metric may be written in closed form as:

$$c_{x,v}(t) = x\exp(tx^T v)\ (\text{or, alternatively } c_{x,x\omega}(t) = x\exp(t\omega).) \tag{82}$$

Also, the Riemannian gradient $\nabla_x V$ with respect to the metrics (80) has the expression:

$$\nabla_x V = \frac{1}{2}\left(\partial_x V - x\partial_x^T V x\right). \tag{83}$$

D. Dynamics on the Grassmann manifold

A Grassmann manifold $\mathrm{Gr}(n,p)$ is a set of subspaces of $\mathbb{R}^{n\times n}$ spanned by p independent vectors. A representation of any of such subspace may be assumed as the equivalence class $[x] = \{x\rho | x \in \mathrm{St}(n,p),\ \rho \in \mathbb{R}^{p\times p},\ \rho^T\rho = e_p\}$, which is the representation used in [12]. In practice, an element $[x]$ of the Grassmann manifold $\mathrm{Gr}(n,p)$ is represented by a matrix in $\mathrm{St}(n,p)$ whose columns span the subspace $[x]$. According to the above representation in terms of a Stiefel matrix, it is possible to express all the quantities of interest in closed form.

For every element $[x] \in \mathrm{Gr}(n,p)$, the tangent space may be represented as:

$$T_{[x]}\mathrm{Gr}(n,p) = \{v \in \mathbb{R}^{n\times p} | x^T v = 0\}, \tag{84}$$

with metric:

$$\langle u, v\rangle_{[x]} = \mathrm{tr}(u^T v),\ \forall u, v \in T_{[x]}\mathrm{Gr}(n,p). \tag{85}$$

A geodesic arc on the Grassmann manifold emanating from $[x] \in \mathrm{Gr}(n,p)$ with velocity $v \in T_{[x]}\mathrm{Gr}(n,p)$ may be written as:

$$c_{[x],v}(t) = [x\beta\ \alpha]\begin{bmatrix}\cos(\sigma t)\\ \sin(\sigma t)\end{bmatrix}\beta^T, \tag{86}$$

where $\alpha\sigma\beta^T$ is the compact singular value decomposition of the matrix v. The Riemannian gradient $\nabla_{[x]}V$ may be calculated as well and reads:

$$\nabla_{[x]}V = (e_n - xx^T)\partial_x V. \tag{87}$$

E. Dynamics on the generalized flag manifold

The generalized flag manifold was introduced in [46] and constitutes a generalization of both the Stiefel and the Grassmann manifolds. A typical application is independent subspace analysis, where signals within any group are allowed to be dependent of each other but signals belonging to different groups are statistically independent.

Take an increasing sequence of subspaces

$$\Psi_1 \subset \Psi_2 \subset \ldots \subset \Psi_r \subset \mathbb{R}^n, \tag{88}$$

with $1 \leq r \leq n$ and the vector space

$$\Psi \stackrel{\text{def}}{=} \Psi_1 \oplus \Psi_2 \oplus \ldots \oplus \Psi_r \subset \mathbb{R}^n, \tag{89}$$

with:

$$\dim \Psi_i = n_i, \ n_1 \leq n_2 \leq \ldots \leq n_r, \tag{90}$$

$$\dim \Psi = \sum_{i=1}^{r} n_i = p \leq n. \tag{91}$$

The set of all such vector spaces is termed flag manifold and is denoted by $\mathrm{Fl}(n, n_1, n_2, \ldots, n_r)$. In the case that all $n_i = 1$, the flag manifold is locally isomorphic to the Stiefel manifold $\mathrm{St}(n, p)$, while in the case that $r = 1$, the flag manifold reduces to the Grassmann manifold $\mathrm{Gr}(n, p)$. The flag manifold is a compact manifold.

A point in the generalized flag manifold $[x] \in \mathrm{Fl}(n, n_1, n_2, \ldots, n_r)$ may be represented by matrices $x \in \mathrm{St}(n, p)$ that further obey the following rule: all the matrices $x\mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_r)$ with $\rho_i \in \mathbb{R}^{n_i \times n_i}$, $\rho_i^T \rho_i = e_{n_i}$, $i = 1, \ 2, \ \ldots, \ r$, represent the same point on the generalized flag manifold as the matrix x.

It is convenient to partition the matrix representing a given point $[x] \in \mathrm{Fl}(n, n_1, n_2, \ldots, n_r)$ as follows:

$$x = [x_{(1)} \ x_{(2)} \ \ldots \ x_{(r)}], \ x_{(i)} \in \mathbb{R}^{n \times d_i}, \ i = 1, \ 2, \ \ldots, \ r. \tag{92}$$

In practice, any submatrix $x_{(i)}$ represents a Grassmann manifold $\mathrm{Gr}(n, n_i)$ with the extra constraint that the global matrix x must be $\mathrm{St}(n, p)$. Any tangent vector $v \in T_{[x]}\mathrm{Fl}(n, n_1, n_2, \ldots, n_r)$ obeys a similar partition. The tangent spaces of the generalized flag manifold have the following structure:

$$\begin{gathered} T_{[x]}\mathrm{Fl}(n, n_1, n_2, \ldots, n_r) = \\ \{v = [v_{(1)} \ v_{(2)} \ \ldots \ v_{(r)}] \in \mathbb{R}^{n \times p} \text{ such that} \\ v_{(i)}^T x_{(i)} = 0, i = 1, \ 2, \ \ldots, \ r, v^T x + x^T v = 0\}. \end{gathered} \tag{93}$$

A metric for the generalized flag manifold is the canonical metric of the Stiefel manifold.

The Riemannian gradient $\nabla_x V$ of a potential energy $V : \mathrm{Fl}(n, n_1, n_2, \ldots, n_r) \to \mathbb{R}$, being a tangent vector, obeys the same partition rule (92), where:

$$(\nabla_x V)_{(i)} = \left(e_n - x_{(i)} x_{(i)}^T\right) \partial_{x_{(i)}} V - \sum_{j \neq i} x_{(j)} \partial_{x_{(j)}}^T V x_{(i)}. \tag{94}$$

The expression of the geodesic may be taken as the one already seen for the Stiefel manifold as well as the formula derived in [45]:

$$c_{[x],v}(t) = \exp(t(\tilde{v}x^T - x\tilde{v}^T))x, \ \tilde{v} \stackrel{\text{def}}{=} \left(e_n - \frac{1}{2} x x^T\right) v. \tag{95}$$

F. Dynamics on the group of symmetric positive-definite matrices

Symmetric positive-definite matrices find a wide range of applications in machine learning. For instance, symmetric positive-definite matrices are applied in low-rank approximation of correlation matrices [26], in the analysis of deformation [53], [56], in pattern recognition, in automatic and intelligent control [8], in the estimation of the power spectrum of random processes [59], in cognitive computation [23] and in the modeling of the functioning of the hypercolumns of the cortical visual area V1 by structure tensors [9].

The manifold of symmetric positive definite matrices is defined as:

$$\mathrm{S}^+(n) = \{x \in \mathbb{R}^{n \times n} | x^T - x = 0, \ x > 0\}. \tag{96}$$

The tangent spaces have structure:

$$T_x \mathrm{S}^+(n) = \{v \in \mathbb{R}^{n \times n} | v^T - v = 0\}, \tag{97}$$

and the canonical metric adopted for the manifold of symmetric positive definite matrices is:

$$\langle u, v \rangle_x = \mathrm{tr}(u x^{-1} v x^{-1}). \tag{98}$$

In order to find the Christoffel matrix function $\Gamma_x(\cdot, \cdot)$ and the associated geodesic equation, the following variational problem needs to be addressed:

$$\int_0^1 \delta \mathrm{tr}(\dot{x} x^{-1} \dot{x} x^{-1}) dt = 0, \tag{99}$$

which is equivalent to:

$$\int_0^1 \mathrm{tr} \left(\frac{d}{dt}(\delta x) x^{-1} \dot{x} x^{-1} + \delta(x^{-1}) \dot{x} x^{-1} \dot{x} \right) dt = 0.$$

From the identity $x x^{-1} = e$, it follows $\delta(x^{-1}) = -x^{-1}(\delta x) x^{-1}$. Substituting the expression of the variation $\delta(x^{-1})$ into the above equation and integrating the first term by parts, gives:

$$\int_0^1 \mathrm{tr} \left(\left(\frac{d}{dt}(x^{-1} \dot{x} x^{-1}) + x^{-1} \dot{x} x^{-1} \dot{x} x^{-1} \right) \delta x \right) dt = 0.$$

The expression in parentheses that multiply the arbitrary symmetric variation δx is symmetric too, therefore the above equation is satisfied if and only if:

$$\frac{d}{dt}(x^{-1} \dot{x} x^{-1}) + (x^{-1} \dot{x})^2 x^{-1} = 0.$$

By computing the derivative with respect to the parameter t and by recalling that $\frac{d}{dt} x^{-1} = -x^{-1} \dot{x} x^{-1}$, the following Christoffel matrix-function arises:

$$\Gamma_x(v, v) = -v x^{-1} v, \ v \in T_x \mathrm{S}^+(n) \tag{100}$$

whose associated geodesic curve reads [43]:

$$c_{x,v}(t) = x^{\frac{1}{2}} \exp(t x^{-\frac{1}{2}} v x^{-\frac{1}{2}}) x^{\frac{1}{2}}. \tag{101}$$

The above expression requires the evaluation of square root of a symmetric positive-definite matrix which exists surely. By recalling that $x^{-1} \exp(v) x = \exp(x^{-1} v x)$, the above expression simplifies into, e.g., $c_{x,v}(t) = x \exp(t x^{-1} v)$.

The Riemannian gradient $\nabla_x V$ of the potential energy function $V : \mathrm{S}^+(n) \to \mathbb{R}$ may be calculated as the unique vector in $T_x \mathrm{S}^+(n)$ that satisfies the following equation:

$$\mathrm{tr}\left(v \partial_x V\right) = \mathrm{tr}\left(v x^{-1} (\nabla_x V) x^{-1}\right), \ \forall v \in T_x \mathrm{S}^+(n). \tag{102}$$

The solution of the above equation satisfies:

$$\begin{aligned} &\partial_x V - x^{-1} (\nabla_x V) x^{-1} = \omega, \\ &\omega = \frac{1}{2} \left(\partial_x V - \partial_x^T V \right), \end{aligned}$$

hence the expression of the Riemannian gradient follows:

$$\nabla_x V = \frac{1}{2} x \left(\partial_x V + \partial_x^T V \right) x. \tag{103}$$

G. Dynamics on the real symplectic group

The real symplectic group is defined as follows:

$$\mathrm{Sp}(2n, \mathbb{R}) = \{x \in \mathbb{R}^{2n \times 2n} | x^T q x = q\}, \ q = \begin{bmatrix} 0_n & e_n \\ -e_n & 0_n \end{bmatrix}, \tag{104}$$

where symbol e_n denotes again the $n \times n$ identity matrix, while symbol 0_n denotes a whole-zero $n \times n$ matrix. The skew-symmetric matrix q enjoys the following properties: $q^2 = -e_{2n}$, $q^{-1} = q^T = -q$.

In the study of optical systems in ophthalmology, it is assumed that the first-order optical nature of an optical system is completely

described by a real matrix $\tau \in \mathbb{R}^{5\times 5}$ termed 'transference' of the optical system [29], [30]. In the most general case, a transference matrix has the form:

$$\tau \stackrel{\text{def}}{=} \left[\begin{array}{cc} s & \ell \\ 0 & 1 \end{array} \right], \tag{105}$$

where $\ell \in \mathbb{R}^4$ and the submatrix s is symplectic, namely, it belongs to the set $\mathrm{Sp}(4,\mathbb{R})$. When an optical system is centered, it holds $\ell = 0$, therefore, the system may be represented by the symplectic submatrix only. The real symplectic groups play an important role in quantum mechanics as well [27]. An important application is quantum computing. Typical gates that quantum computers make use of are Hadamard gates, phase gates, CNOT gates and Pauli gates, and the basic information unit is the "qubit" [5]. In addition, the real symplectic group plays a role in the study of Monge-Ampére equations [39].

The space $\mathrm{Sp}(2n,\mathbb{R})$ is a curved smooth manifold that may also be endowed with smooth algebraic-group operations in a manner that is compatible with the manifold structure. Therefore, the space $\mathrm{Sp}(2n,\mathbb{R})$ has the structure of a Lie group. In particular, standard matrix multiplication and inverse work as algebraic group operations. The identity element of the group $\mathrm{Sp}(2n,\mathbb{R})$ is clearly the matrix e_{2n}.

The tangent space $T_x\mathrm{Sp}(2n,\mathbb{R})$ has the structure:

$$T_x\mathrm{Sp}(2n,\mathbb{R}) = \{v \in \mathbb{R}^{2n\times 2n} | v^T qx + x^T qv = 0_{2n}\}. \tag{106}$$

In particular, the tangent space at the identity of the Lie group, namely the Lie algebra $\mathfrak{sp}(2n,\mathbb{R})$, has the structure:

$$\mathfrak{sp}(2n,\mathbb{R}) = \{h \in \mathbb{R}^{2n\times 2n} | h^T q + qh = 0\}, \tag{107}$$

namely, it coincides with the space of $2n\times 2n$ Hamiltonian matrices. The tangent space, the Lie algebra and the normal space associated to the real symplectic group may be characterized as follows:

$$\left\{ \begin{array}{l} T_x\mathrm{Sp}(2n,\mathbb{R}) = \{v = xqs | s \in \mathbb{R}^{2n\times 2n}, s^T = s\}, \\ \mathfrak{sp}(2n,\mathbb{R}) = \{h = qs | s \in \mathbb{R}^{2n\times 2n}, s^T = s\}, \\ N_x\mathrm{Sp}(2n,\mathbb{R}) = \{n = qx\omega | \omega \in \mathbb{R}^{2n\times 2n}, \omega^T = -\omega\}. \end{array} \right. \tag{108}$$

The real symplectic group appears to be far less studied than the previously considered manifolds. A result that may be of use in order to study extended Hamiltonian systems on $\mathrm{Sp}(2n,\mathbb{R})$ is adapted from [6]. Let $\sigma : \mathfrak{sp}(2n,\mathbb{R}) \to \mathfrak{sp}(2n,\mathbb{R})$ be a symmetric positive-definite operator with respect to the Euclidean inner product on the space $\mathfrak{sp}(2n,\mathbb{R})$. The minimizing curve of the integral:

$$\int_0^1 \mathrm{tr}((x^{-1}\dot{x})^T \sigma(x^{-1}\dot{x}))dt, \tag{109}$$

over all curves $x(t) \in \mathrm{Sp}(2n,\mathbb{R})$ with $t \in [0\ 1]$ and with fixed endpoints $x(0) = x_1 \in \mathrm{Sp}(2n,\mathbb{R})$ and $x(1) = x_2 \in \mathrm{Sp}(2n,\mathbb{R})$ is the solution of the system:

$$\dot{x} = xh,\ \dot{m} = \sigma^T(h)m - m\sigma^T(h),\ h = \sigma^{-1}(m), \tag{110}$$

where symbol σ^{-1} denotes the inverse of the operator σ.

Closed-form solutions of the above system are presently unknown. The simplest choice for the symmetric positive-definite operator σ is $\sigma(h) = h$, which corresponds to a metric for the real symplectic group $\mathrm{Sp}(2n,\mathbb{R})$ given by:

$$\langle u, v\rangle_x = \mathrm{tr}((x^{-1}u)^T(x^{-1}v)),\ \forall u, v \in T_x\mathrm{Sp}(2n,\mathbb{R}). \tag{111}$$

Such a metric basically corresponds to the metric which leads to the 'natural gradient' on the space of real invertible matrices $\mathrm{Gl}(n,\mathbb{R})$ studied in [3]. The above choice for the operator σ implies that the corresponding geodesic curve on the real symplectic group satisfies the equations:

$$\dot{h} = h^T h - hh^T,\ h = x^{-1}\dot{x}. \tag{112}$$

The Christoffel form associated to the Riemannian metric (111) is, thus, found to be:

$$\Gamma_x(v,v) = -vx^{-1}v + xv^T qxqx^{-1}v - vv^T qxq,\ v \in T_x\mathrm{Sp}(2n,\mathbb{R}). \tag{113}$$

The Riemannian gradient $\nabla_x V$ of a potential energy function $V : \mathrm{Sp}(2n,\mathbb{R}) \to \mathbb{R}$ may be calculated as the unique vector in $T_x\mathrm{Sp}(2n,\mathbb{R})$ that satisfies the metric-compatibility equation:

$$\mathrm{tr}\left(v^T \partial_x V\right) = \mathrm{tr}\left((x^{-1}v)^T(x^{-1}\nabla_x V)\right),\ \forall v \in T_x\mathrm{Sp}(2n,\mathbb{R}). \tag{114}$$

Recalling the structures of the tangent and normal spaces to the real symplectic group at a point, the solution of the above equation is such that:

$$\begin{aligned} \nabla_x V &= xq(\omega - x^{-1}q\partial_x V), \\ \omega &= \frac{1}{2}x^{-1}q\partial_x V + \frac{1}{2}\partial_x^T V xq, \end{aligned}$$

from which the expression of the sought-for Riemannian gradient

$$\nabla_x V = \frac{1}{2}xq\left(\partial_x^T V xq - qx^T \partial_x V\right) \tag{115}$$

immediately follows.

IV. Conclusion

The present contribution aims at introducing a general framework to develop a theory of learning on differentiable manifolds by extended Hamiltonian stationary-action principle.

The first part of the paper presents the derivation of general equations of learning on Riemannian manifolds as well as the energy balance equation for a general extended Hamiltonian system. The first part of the paper also discusses the relationship between the presented learning theory (namely, dynamical learning by second-order differential equations on manifolds) and the classical learning theory based on gradient-steepest-descent optimization. An interesting finding is the resemblance of dynamical learning with classical learning with 'momentum term'. A convergence analysis is carried out in terms of Lyapunov function and in terms of fine-convergence analysis in the proximity of a stationary point of the learning-goal function, represented by a potential energy function for the dynamical system. The fine-convergence analysis shows that second-order learning may improve the convergence ability of first-order learning if certain conditions are met. Such conditions may be met by properly selecting the damping coefficient of the dynamical system.

The second part of the paper discusses several cases of learning by extended Hamiltonian systems on manifolds of interest in the scientific literature (namely, the Stiefel manifold, the manifold of multidimensional rotations, the Grassmann manifold, the manifold of symmetric positive definite matrices, the generalized flag manifold and the manifold of real symplectic matrices). The purpose of the calculations is to show how to derive the Christoffel form, the Riemannian gradient and the geodesic expressions by working in intrinsic coordinates only by means of the variational principles stated in the first part of the paper. The same techniques may be applied to other manifolds of possible interest in the scientific literature that have not been taken explicitly into account in the present manuscript. The obtained results allow formulating the dynamical-learning equations for the manifolds of interest and are prodromic to their numerical implementation.

The present contribution focused on real-valued matrix-type manifolds. Complex-valued manifolds are of interest too in the literature of neural learning and adaptive signal/data processing. The extension of the introduced extended Hamiltonian learning theory to the complex domain does not appear particularly difficult, therefore it has not

been treated here. (An early attempt to stretch-out the extended Hamiltonian learning on the unit hypersphere S^{n-1} to the complex domain was published in [24]. A more recent study tailored to the case of the low-dimensional special unitary group $\mathrm{SU}(3)$ appeared in [21].)

The formulation of the extended Hamiltonian learning equations on differentiable manifolds requires instruments from differential geometry. Likewise, the numerical implementation of the extended Hamiltonian systems on a computation platform requires instruments from *geometric integration* [28].

To render the problem that arises about the numerical implementation of the dynamical learning equations on manifolds, it is worth examining again the explanatory example of section II. Suppose a learning problem is formulated in terms of the first-order differential equation $\dot{x} = F(x)$ on the manifold $\mathrm{SO}(2)$, with $F : x \in \mathrm{SO}(2) \mapsto F(x) \in T_x\mathrm{SO}(2)$. Recall that in the context of implementation of learning equations only intrinsic coordinates are made use of because the use of extrinsic coordinates is unpractical. In intrinsic coordinates, according to the notation introduced in (4), the differential equation writes:

$$\frac{d}{dt}\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} F_{11}(x_{11},x_{12},x_{21},x_{22}) & F_{12}(x_{11},x_{12},x_{21},x_{22}) \\ F_{21}(x_{11},x_{12},x_{21},x_{22}) & F_{22}(x_{11},x_{12},x_{21},x_{22}) \end{bmatrix}.$$

The standard stepping techniques of numerical calculus, such as Euler methods, may not be applied to solve the above differential equation. In fact, it is clear that a numerical method such as, for example:

$$x_{ij}(t+\Delta t) = x_{ij}(t) + \epsilon F_{ij}(x_{11}(t), x_{12}(t), x_{21}(t), x_{22}(t))$$

does not take into account the constraints (5), namely, it generates a trajectory $x(t)$ in the ambient space $\mathbb{R}^{2\times 2}$ rather than in the feasible space $\mathrm{SO}(2)$. Namely, starting from a point $x(t) \in \mathrm{SO}(2)$, it would produce a new point $x(t+\Delta t) \notin \mathrm{SO}(2)$. The reason of such behavior is that the Euler numerical integration techniques insist on the flat space $\mathbb{R}^n$ and do not cope (in general) with curved manifolds. For a general parameter manifold M, the above stepping method should be replaced with a numerical method that is capable of coping with the curved nature of the parameter space, which may be written:

$$x(t+\Delta t) = \mathcal{H}(x(t), \epsilon),$$

where the operator $\mathcal{H} : M \times \mathbb{R} \to M$ is designed in order to ensure that starting from a point $x(t) \in M$, it will produce a new point $x(t+\Delta t) \in M$. Geometric integration is a branch of numerical calculus whose goal is to numerically solve systems of differential equations on differentiable manifolds [28].

The same kind of question arises about the numerical integration of the the dynamical system (21). As the formulation of suitable numerical integration methods for the discussed extended Hamiltonian systems is a quite extended topics (for an example tailored to the case of Stiefel-manifold-learning, see [19]), it has not been treated here and will be treated on a research paper which is currently in preparation.

V. Acknowledgments

The Author wishes to gratefully thank the Associate Editor as well as the anonymous Reviewers for their detailed and careful comments, which helped improve the quality of presentation and the thoroughness of the technical content of the manuscript.

References

[1] P.-A. Absil, R. Mahony and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008
[2] C. Aluffi-Pentini, V. Parisi and F. Zirilli, *Global optimization and stochastic differential equations*, Journal of Optimization Theory and Applications, Vol. 47, pp. 1 – 16, 1985
[3] S.-i. Amari, *Natural gradient works efficiently in learning*, Neural Computation, Vol. 10, 251 – 276, 1998
[4] S.-i. Amari, *Natural gradient learning for over- and under-complete bases in ICA*, Neural Computation, Vol. 11, pp. 1875 – 1883, 1999
[5] S. D. Bartlett, B. Sanders, S. Braunstein and K. Nemoto, *Efficient classical simulation of continuous variable quantum information processes*, Physical Review Letters, Vol. 88, 097904/1-4, 2002
[6] A.M. Bloch, P.E. Crouch, J.E. Marsden and A.K. Sayal, *Optimal control and geodesics on quadratic matrix Lie groups*, Foundations of Computational Mathematics, Vol. 8, pp. 469 – 500, 2008
[7] E. Celledoni and S. Fiori, *Neural learning by geometric integration of reduced 'rigid-body' equations*, Journal of Computational and Applied Mathematics (JCAM), Vol. 172, No. 2, pp. 247 – 269, December 2004
[8] Y. Chen and J.E. McInroy, *Estimation of symmetric positive-definite matrices from imperfect measurements*, IEEE Trans. on Automatic Control, Vol. 47, No. 10, pp. 1721 – 1725, October 2002
[9] P. Chossat and O. Faugeras, *Hyperbolic planforms in relation to visual edges and textures perception*, PLoS Computational Biology. DOI: 10.1371/journal.pcbi.1000625, 2009
[10] A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*, John Wiley Ltd, 1993
[11] K.I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks: Theory and Applications*, Wiley-Interscience, 1996
[12] A. Edelman, T.A. Arias and S.T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis Applications, Vol. 20, No. 2, pp. 303 – 353, 1998
[13] L. Eldén and H. Park, *A Procrustes problem on the Stiefel manifold*, Numerical Mathematics, Vol. 82, pp. 599 – 619, 1999
[14] S. Fiori, *A theory for learning based on rigid bodies dynamics*, IEEE Trans. on Neural Networks, Vol. 13, No. 3, pp. 521 – 531, May 2002
[15] S. Fiori, *Unsupervised neural learning on Lie groups*, International Journal of Neural Systems, Vol. 12, No.s 3 & 4, pp. 219 – 246, 2002
[16] S. Fiori, *Neural minor component analysis approach to robust constrained beamforming*, IEE Proceedings - Vision, Image and Signal Processing, Vol. 150, No. 4, pp. 205 – 218, August 2003
[17] S. Fiori, *A fast fixed-point neural blind deconvolution algorithm*, IEEE Trans. on Neural Networks, Vol. 15, No. 2, pp. 455 – 459, March 2004
[18] S. Fiori, *Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial*, Journal of Machine Learning Research, Vol. 6, pp. 743 – 781, May 2005
[19] S. Fiori, *Formulation and integration of learning differential equations on the Stiefel manifold*, IEEE Trans. on Neural Networks, Vol. 16, No. 6, pp. 1697 – 1701, November 2005
[20] S. Fiori, *Geodesic-based and projection-based neural blind deconvolution algorithms*, Signal Processing, Vol. 88, No. 3, pp. 521 – 538, March 2008
[21] S. Fiori, *A study on neural learning on manifold foliations: The case of the Lie Group SU(3)*, Neural Computation, Vol. 20, No. 4, pp. 1091 – 1117, April 2008
[22] S. Fiori, *Learning by natural gradient on noncompact matrix-type pseudo-Riemannian manifolds*, IEEE Trans. on Neural Networks, Vol. 21, No. 5, pp. 841 – 852, May 2010
[23] S. Fiori, *Learning the Fréchet mean over the manifold of symmetric positive-definite matrices*, Cognitive Computation (Springer). Vol. 1, No. 4, pp. 279 – 291, December 2009
[24] S. Fiori and P. Burrascano, *One-unit 'rigid-bodies' learning rule for principal/independent component analysis with application to ECT-NDE signal processing*, Neurocomputing, Vol. 56, pp. 233 – 255, January 2004
[25] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Dover Publications, 2000
[26] I. Grubišić and R. Pietersz, *Efficient rank reduction of correlation matrices*, Preprint of the Utrecht University, January 2005
[27] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics*, Cambridge University Press, 1984
[28] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics, 2nd Edition, 2006

[29] W.F. Harris, *Paraxial ray tracing through noncoaxial astigmatic optical systems, and a* 5×5 *augmented system matrix*, Optometry and Vision Science, Vol. 71, No. 4, pp. 282 – 285, 1994
[30] W.F. Harris, *The average eye*, Ophthalmic and Physiological Optics, Vol. 24, pp. 580 – 585, 2004
[31] K.A. Hoffman, *Methods for determining stability in continuum elastic-rod models of DNA*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 362, No. 1820, pp. 1301 – 1315, July 2004
[32] R. Hunger, P. de Kerret and M. Joham, *An algorithm for maximizing a quotient of two hermitian form determinants with different exponents*, Proceeding of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010, Dallas (TX, USA), March 2010), pp. 3346 – 3349, 2010
[33] M. Joho and K. Rahbar, *Joint diagonalization of correlation matrices by using Newton methods with applications to blind signal separation*, Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop, pp. 403 – 407, 2002
[34] K. Kreutz-Delgado and B.D. Rao, *Sparse basis selection, ICA, and majorization: Towards a unified perspective*, Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 1999, Phoenix (AZ, USA), March 1999), Vol. 2, pp. 1081 – 1084, 1999
[35] D. Isvoranu and C. Udrişte, *Fluid flow versus geometric dynamics*, 5^{th} Conference on Differential Geometry (Mangalia (Romania), August - September 2005), BSG Proceedings 13, pp. 70 – 82, Geometry Balkan Press, 2006
[36] C. Lanczos, *The Variational Principles of Mechanics*, Dover Books on Physics and Chemistry, 4th Edition, 1986
[37] D.G. Luenberger, *The gradient projection method along geodesics*, Management Science, Vol. 18, No. 11, pp. 620 – 631, July 1972
[38] T.-W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic Publishers, September 1998
[39] V.V. Lychagin, V.N. Rubtsov and I.V. Chekalov, *A classification of Monge-Ampére equations*, Annales Scientifiques de l'École Normale Supérieure, Vol. 26, No. 4, pp. 281 – 308, 1993
[40] C.S. MacInnes and R.J. Vaccaro, *Tracking direction-of-arrival with invariant subspace updating*, Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 1996), pp. 2896 – 2899, 1996
[41] R.S. Manning and G.B. Bulman, *Stability of an elastic rod buckling into a soft wall*, Proc. of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 461, No. 2060, pp. 2423 – 2450, August 2005
[42] J. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, Texts in Applied Mathematics 17, Springer-Verlag, 2nd Edition, 1999
[43] M. Moakher, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM Journal of Matrix Analysis and Applicatrions, Vol. 26, No. 3, pp. 735–747, 2005
[44] E. Moreau and J.C. Pesquet, *Independence/decorrelation measures with application to optimized orthonormal representations*, Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 1997, Munich (Germany), April 1997), Vol. 5, pp. 3425 – 3428, 1997
[45] Y. Nishimori and S. Akaho, *Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold*, Neurocomputing (Special issue on "Geometrical methods in neural networks and learning", S. Fiori and S.-i. Amari, Ed.s), Vol. 67, pp. 106 – 135, 2005
[46] Y. Nishimori, S. Akaho and M.D. Plumbley, *Riemannian optimization method on the flag manifold for independent subspace analysis*, Proc. of the 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06), Lecture notes in computer science, Vol. 3889, pp. 295 – 302, Springer, Berlin, 2006
[47] P. Pajunen and M. Girolami, *Implementing decisions in binary decision trees using independent component analysis*, Proc. of the International Workshop on Independent Component Analysis and Blind Signal Separation, pp. 477 – 481 (June 19-22, 2000, Helsinki, Finland), 2000
[48] F. Palmieri, J. Zhu and C. Chang, *Anti-Hebbian learning in topologically constrained linear networks: A tutorial*, IEEE Trans. on Neural Networks, Vol. 4, No. 5, pp. 748 – 761, 1993
[49] A.J. Pellionisz, *Coordination: A vector-matrix description of transformations of overcomplete CNS coordinates and a tensorial solution using the Moore-Penrose generalized inverse*, Journal of Theoretical Biology, Vol. 110, pp. 353 – 375, 1984
[50] A.J. Pellionisz and R. Llinás, *Tensorial approach to the geometry of brain function. Cerebellar coordination via a metric tensor*, Neuroscience, Vol. 5, pp. 1761 – 1770, 1980
[51] M.J. Prentice, *Fitting smooth paths to rotation data*, The Journal of the Royal Statistical Society Series C (Applied Statistics), Vol. 36, No. 3, pp. 325 – 331, 1987
[52] N. Qian, *On the momentum term in gradient descent learning algorithms*, Neural Networks, Vol. 12, No. 1, pp. 145 – 151, January 1999
[53] I.U. Rahman, I. Drori, V.C. Stodden, D.L. Donoho and P. Schröder, *Multiscale representations for manifold-valued data*, Multiscale Modeling and Simulation, Vol 4, No. 4, pp 1201 – 1232, 2005
[54] R.P.N. Rao and D.L. Ruderman, *Learning Lie groups for invariant visual perception*, Advances in Neural Information Processing Systems (NIPS) 11, pp. 810 – 816, 1999
[55] Q. RENTMEESTERS, P.-A. ABSIL, P. VAN DOOREN, K. GALLIVAN AND A. SRIVASTAVA, *An efficient particle filtering technique on the Grassmann manifold*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, Dallas (TX, USA), March 14-19, 2010), pp. 3838 – 3841, 2010
[56] J. Salencon, *Handbook of Continuum Mechanics*, Springer-Verlag, Berlin, 2001
[57] O. Shalvi and E. Weinstein, *Super-exponential methods for blind deconvolution*, IEEE Trans. on Information Theory, vol. 39, pp. 504 – 519, March 1993
[58] S. Shorek, *A stationarity principle for non-conservative systems*, Advanced Water Resources, Vol. 7, pp. 85 – 88, June 1984
[59] K. Slavakis, I. Yamada and K. Sakaniwa, *Computation of symmetric positive definite Toeplitz matrices by the hybrid steepest descent method*, Signal Processing, Vol. 83, No. 5, pp. 1135 – 1140, May 2003
[60] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, 2^{nd} Edition, Berkeley, CA: Publish or Perish Press, 1979
[61] A. Srivastava, U. Grenander, G.R. Jensen and M.I. Miller, *Jump-diffusion Markov processes on orthogonal groups for object pose estimation*, Journal of Statistical Planning and Inference, Vol. 103, pp. 15 – 37, 2002
[62] P. TURAGA, S. BISWAS AND R. CHELLAPPA, *The role of geometry in age estimation*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, Dallas (TX, USA), March 14-19, 2010), pp. 946 – 949, 2010
[63] B. Vujanovic, *A variational principle for non-conservative dynamical systems*, Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM - Journal of Applied Mathematics and Mechanics), Vol. 55, pp. 321 – 331, 1975
[64] H. Xiong, M.N.S. Swamy and M.O. Ahmad, *Optimizing the kernel in the empirical feature space*, IEEE Transactions on Neural Networks, Vol. 16, No. 2, pp. 460 – 474, 2005
[65] L. Xu, E. Oja and C.Y. Suen, *Modified Hebbian learning for curve and surface fitting*, Neural Networks, Vol. 5, pp. 441 – 457, 1992
[66] J. Yoo and S. Choi, *Orthogonal nonnegative matrix factorization: multiplicative updates on Stiefel manifolds*, Proc. of the Intelligent Data Engineering and Automated Learning (IDEAL 2008), Springer Berlin/Heidelberg, pp. 140 – 147, 2008
[67] J.M. Zurada, *Introduction to Artificial Neural Systems*, PWS Publishing Company, 1992

Simone Fiori received the Italian Laurea (Dr. Eng.) *cum laude* in Electronic Engineering in July 1996 from the University of Ancona (Italy) and the Ph.D. degree in Electrical Engineering (circuit theory) in March 2000 from the University of Bologna (Italy). He is currently serving as Adjunct Professor at the Faculty of Engineering of the Università Politecnica delle Marche (Ancona, Italy). His research interests include unsupervised learning theory for artificial neural networks, linear and non-linear adaptive discrete-time filter theory, vision and image processing by neural networks, continuous-time and discrete-time circuits for stochastic information processing, geometrical methods for machine learning and signal processing. He is author of more than 145 refereed journal and conference papers. Dr. Fiori was the recipient of the 2001 "E.R. Caianiello Award" for the best Ph.D. dissertation in the artificial neural network field and the 2010 "Rector Award" as a proficient researcher of the Faculty of Engineering of the Università Politecnica delle Marche. He is currently serving as Associate Editor for Neurocomputing (Elsevier), Computational Intelligence and Neuroscience (Hindawi) and Cognitive Computation (Springer).