Neural Learning on the Compact Stiefel Manifold by Reduced `Rigid-Body‘ Equations (and their Geometric Integration)

Simone Fiori

Faculty of Engineering, Perugia University (Italy)

(Web: http://www.unipg.it/~sfr/)

Temporary visitor at Neuroscience Research Institute
(AIST, Tsukuba, Japan)

Overview (1)

- Introduction
  - Neural learning as optimization task
  - Orthonormality constraints
  - Rationale for Stiefel-manifold learning

- ‘Rigid-body’ learning theory
  - Rigid-body learning equations
  - Related studies and allied topics
  - Main properties (equilibria, stability)
  - Illustrative examples
  - Notes on implementation

Overview (2)

- Reduced Rigid-Body Learning
  - Representation theorem
  - Reworking the equations
- Learning by Geometric Integration
  - What is geometric integration ?
  - An illustrative example
  - Proposed integration scheme
- Conclusions

Neural Learning as Optimization

- Learning arising from optimization:
  - Description of network's task:
    - 1) synthesis of learning criterion;
    - 2) individuation of the constraints.
  - Selection/design of a proper optimization method:
    - 1) formulation of differential equations describing network dynamics;
    - 2) description of the invariants accounting for the constrains.
  - Implementation of the learning theory:
    - 1) (numerical) integration of the differential equations;
    - 2) preservation of the constraints described by the invariants.

  - Example: Principal subspace analysis (PSA). Criterion: Rayleigh quotient. Constraints: Orthonormality. Differential equation: Oja's flow. Invariant: Stiefel (Grassman) manifold. Integration: Euler method. Preservation of invariant: ?.

Orthonormality constraints

- Let us consider a one-layer network:

$$\mathbf{y} = S(\mathbf{W}^T\mathbf{x} + \mathbf{b}),\ \mathbf{x} \in \Re^p,\ \mathbf{y} \in \Re^m$$

- Constraints of orthonormality: $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{p,m}$.

- The invariant is the Stiefel manifold: $\mathrm{St}(p, m, \Re)$.

  - Note: $m \le p$. When inequality holds the manifold is connected. When equality holds the manifold is isomorphic to the orthogonal group which has two components; in this case, we work with the orthogonal matrices having determinant +1, i.e. the special orthogonal group (SO).

Rationale for Stiefel Learning - 1

- Many well-known applications:
  - Principal/minor component analysis. Subspace iteration.
  - Independent (over- and under-determined) component analysis.

- Several potential applications:
  - Solid-state physics computations (Edelman, Arias and Smith, 1998)
  - Dynamic texture recognition via ARMA modeling of image streams (Saisan, Doretto, Wu and Soatto, 2001)
  - Transformation-invariant optical character recognition (Sona, Sperduti and Starita, 2000)
  - Optical flow estimation by SVD computation and tracking (Fiori, 2003)

Rationale for Stiefel Learning - 2

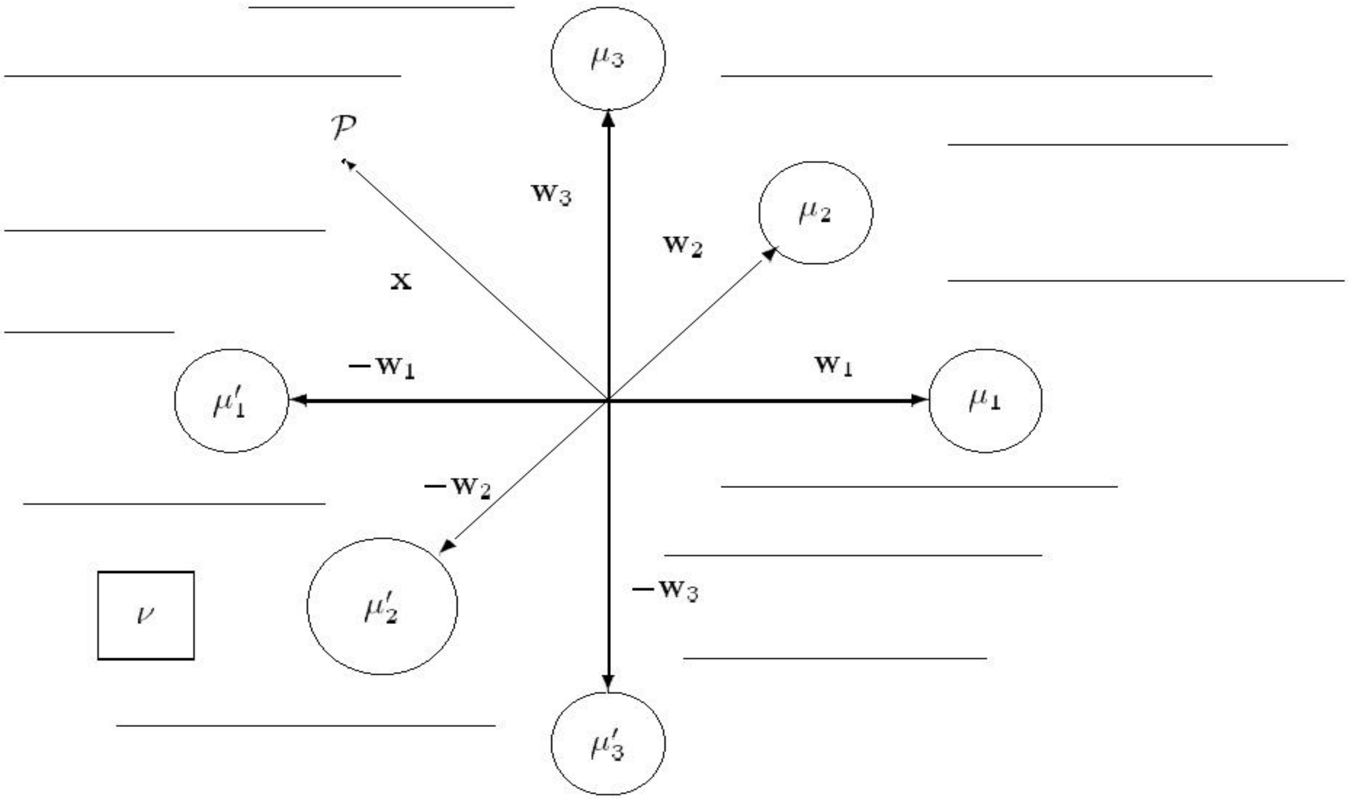
- References for the potential applications:
  - A. Edelman, T. Arias and S.T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis and Applications, Vol. 20, No. 2, pp. 303 – 353, 1998
  - P. Saisan, G. Doretto, Y.N. Wu and S. Soatto, *Dynamic texture recognition*, Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 58 - 63, Dec. 2001
  - D. Sona, A. Sperduti, and A. Starita, *Discriminant Pattern Recognition Using Transformation Invariant Neurons*, Neural Computation, Vol. 12, No. 6, pp. 1355 - 1370, June 2000
  - S. Fiori, *Singular Value Decomposition Learning on Double Stiefel Manifold*, International Journal of Neural Systems. Accepted for publication, 2003

Rationale for Stiefel Learning - 3

- Intrinsic “stability” of the learning method (thanks to the compactness of the parameter space).
  - Compactness ensures the non-divergence of the differential equations.
  - Some applications and thoughts on stability are summarized in:
    - S. Fiori, “A Theory for Learning by Weight Flow on Stiefel-Grassman manifold”, Neural Comp., Vol. 13, No. 7, pp. 1625 – 1647, 2001
    - S. Fiori, “Unsupervised Neural Learning on Lie Group”, Int. J. of Neural Systems, Vol. 12, No. 3-4, pp. 219 – 246, 2002

‘Rigid-Body’ Learning Theory

Natural phenomena may be formalized via optimization.



Rigid-body dynamics theory provides a natural way to formalize general Stiefel manifold learning.

Some Exemplary Contributions and Allied Topics - 1

- Some contributions on natural, second-order, non-gradient methods for neural learning as well as optimization:
  - **C. Aluffi-Pentini, V. Parisi, and F. Zirilli**, Global Optimization and Stochastic Differential Equations, *J. of Optimization Theory and Applications*, Vol. 47, pp. 1 - 16, 1985
  - **S.-i. Amari, H. Park and K. Fukumizu**, Adaptive method of realizing natural gradient learning for multilayer perceptrons, *Neural Computation*, 12, 1399-1409, 2000
  - **R.W. Brockett**, Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems, *Linear Algebra and Its Applications*, Vol. 146, pp. 79 - 91, 1991
  - **S. Hochreiter and M.C. Mozer**, Coulomb classifiers: Reinterpreting SVMs as electrostatic systems, Technical report CU-CS-921-01, Dept. of Computer Science. University of Colorado, May 2001

Some Exemplary Contributions and Allied Topics - 2

- Some contributions on natural, second-order, non-gradient methods for neural learning as well as optimization:

  - **Y. Nishimori**, Learning Algorithm for ICA by Geodesic Flows on Orthogonal Group, *Proc. of the International Joint Conference on Neural Networks (IJCNN'99)*, Vol. 2, pp. 1625 - 1647, 1999
  - **N. Qian**, On the Momentum Term in Gradient Descent Learning Algorithms, *Neural Networks*, Vol. 12, pp. 145 - 151, 1999
  - **K. Zhang and T.J. Sejnowski**, A theory of geometric constraints on neural activity for natural three-dimensional movement, *Journal of Neuroscience*, Vol. 19, No. 8, pp. 3122 -- 3145, 1999

Rigid-body Learning Equations - 1

- Learning equations arising from the study of the dynamics of an abstract rigid body:

$$\frac{d\mathbf{W}}{dt} = \mathbf{BW}\,, \ \mathbf{R} = -n\ \mathbf{BW}\,,$$

$$\frac{d\mathbf{B}}{dt} = \frac{1}{4}\left[(\mathbf{F}+\mathbf{R})\mathbf{W}^{T} - \mathbf{W}(\mathbf{F}+\mathbf{R})^{T}\right],$$

$$\mathbf{F} = -\frac{\partial U(\mathbf{W})}{\partial \mathbf{W}},\ U(\mathbf{W}) = E_{\mathbf{x}}\left[u(\mathbf{W},\mathbf{x},\mathbf{y})\right].$$

- Detailed derivation in: 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

Rigid-Body Learning Equations - 2

- Main properties:
  - The connection matrix **W**(t) belongs to the Stiefel manifold/orthogonal group at any time. The angular-speed matrix **B**(t) belongs to the Lie algebra of the orthogonal group at any time.
  - The fixed points of the learning equations satisfy:

$$\mathbf{BW} = \mathbf{0}\,,\ \mathbf{FW}^T - \mathbf{WF}^T = \mathbf{0}$$

  - The system tends to <u>minimize</u> the potential energy function *U* (*t*).
  - The system is second-order in time (this allows better control of the dynamics for e.g. local extremes avoidance).
  - The equilibria are asymptotically stable.

Rigid-Body Learning Equations - 3

- Proof of stability through Lyapunov function.
  - Let us define the linear velocity and network’s kinetic energy:

$$\mathbf{V}(t) = \mathbf{B}(t)\mathbf{W}(t)\,,\; K(t) = \frac{1}{2} tr[\mathbf{V}^T(t)\mathbf{V}(t)]$$

  - The energy balance equation writes:

$$K(t) - K(0) = -[U(t) - U(0)] - 2n\int_0^t K(t)dt$$

  - If U_* denotes the minimum of U over the manifold (it exists because of regularity and compactenss), the Hamiltonian $H(t) = K(t) + U(t) - U_*$ is – by construction – positive and enjoys the property:

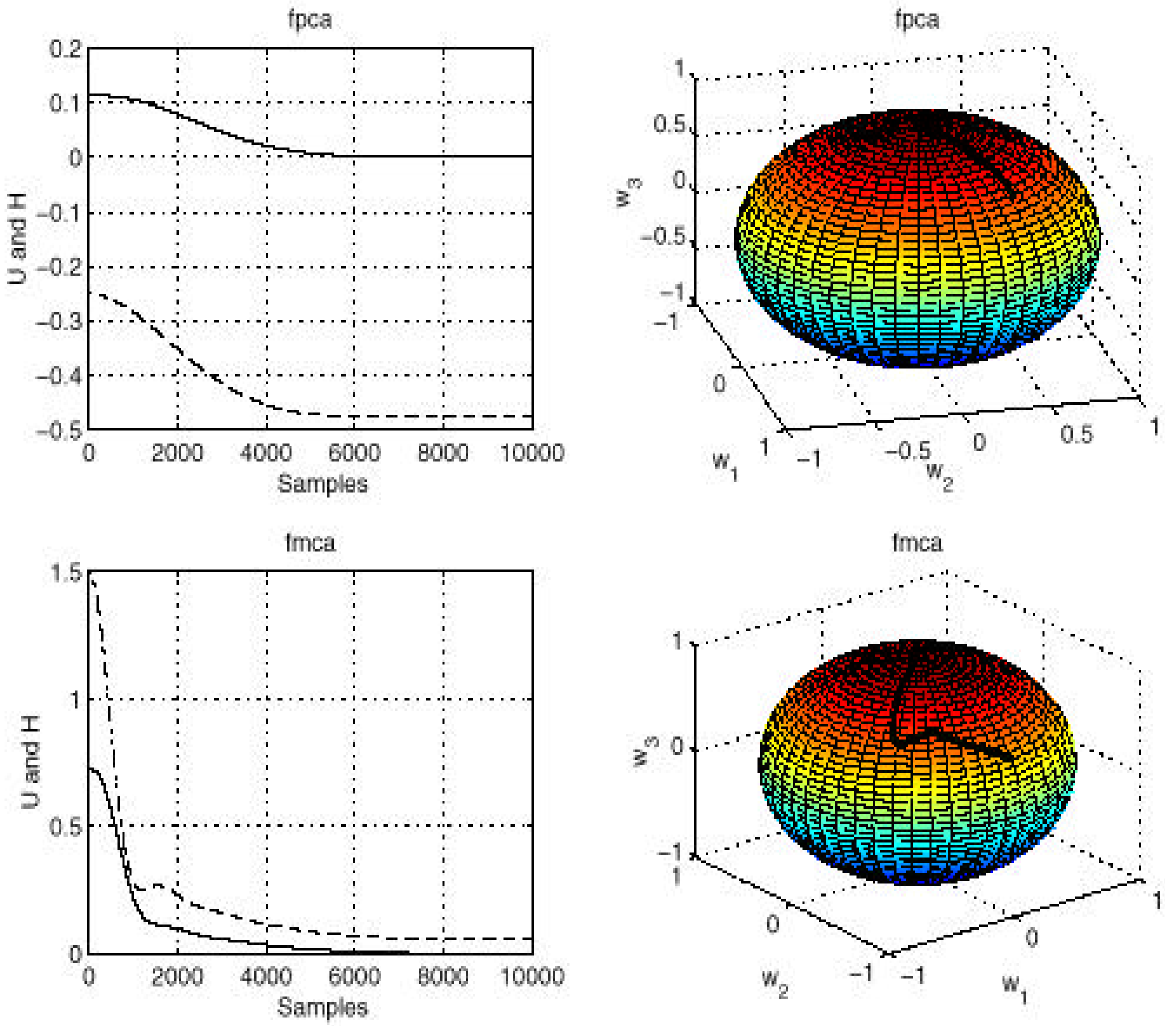
$$H(t) = H(0) - 2n\int_0^t K(t)dt$$

  - Thus its time-derivative is $\dot{H}(t) = -2nK(t) \le 0$. The equality hold at equilibrium.

Illustrative examples - 1

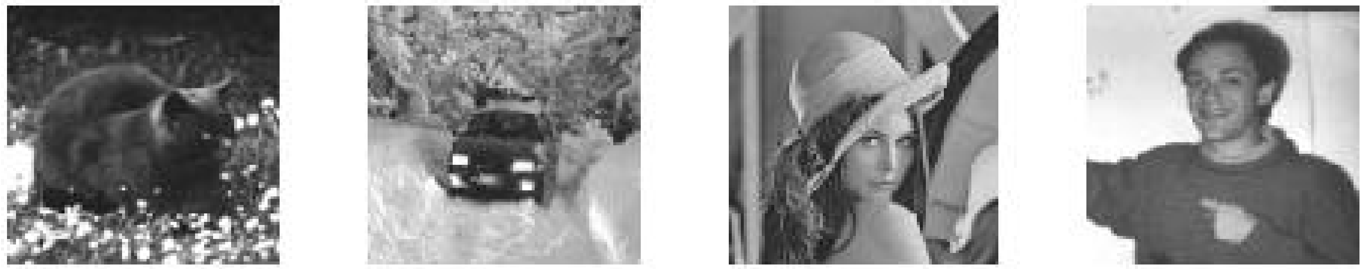
- Case of one-unit first (top)/last (bottom) principal component analysis.
- The total energy (solid-line) H(t) converges to zero (the system looses energy because of friction).
- The potential energy function is minimized.

$$U(\mathbf{w}) = \pm E[y^2]$$

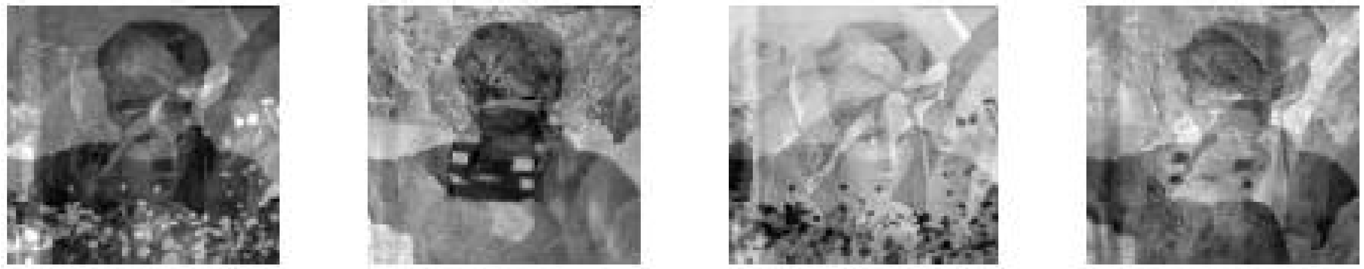


Illustrative examples - 2

- Case of one-unit kurtosis-based independent component analysis.



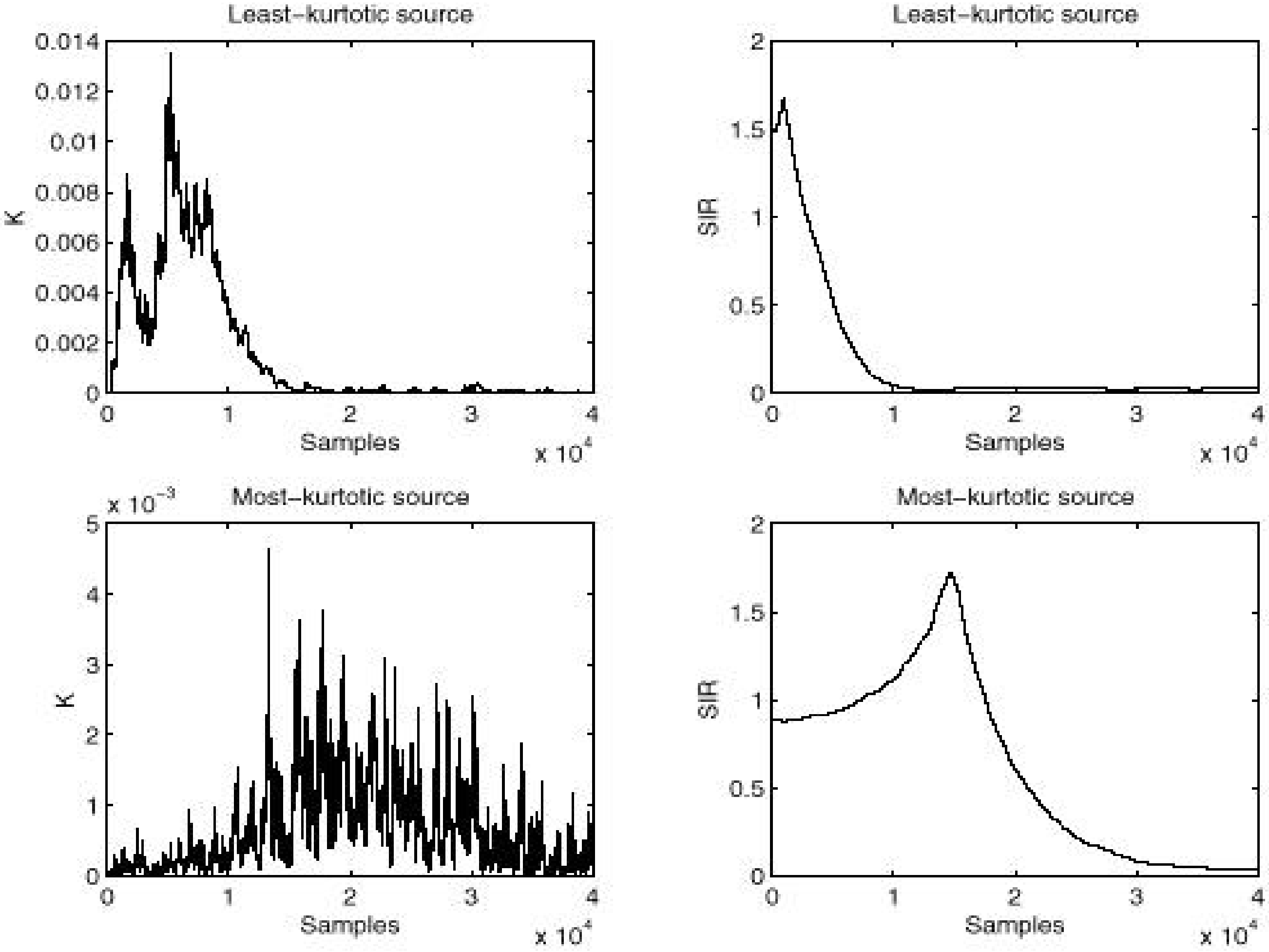
Original sources (top) and mixed images (bottom).



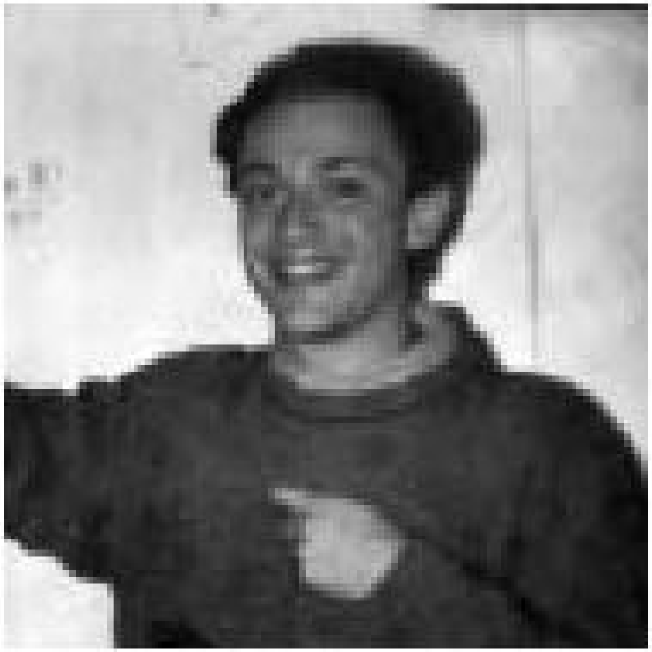
Illustrative examples - 3

- Neuron's kinetic energy <u>and</u> signal-to-interference residual <u>during learning</u>.

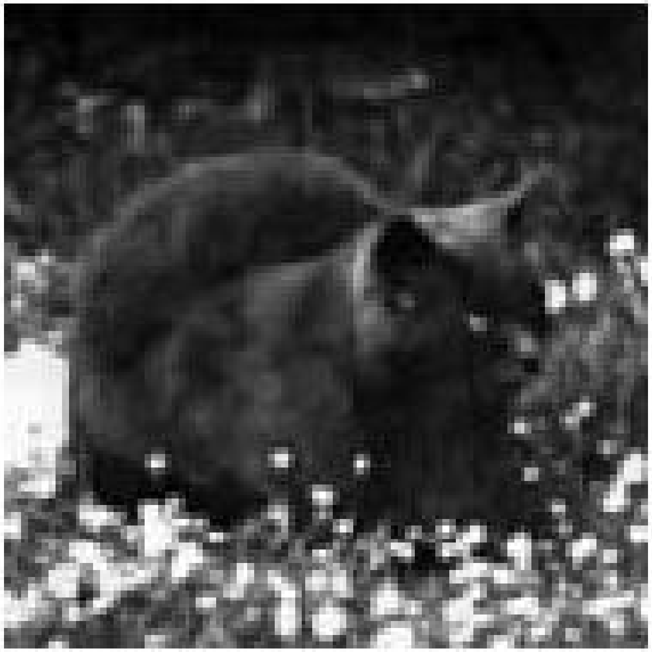
$$U(\mathbf{w}) = \pm E[y^4]$$



Illustrative examples - 4



Recovered least–kurtotic



Recovered most–kurtotic

- Neuron’s output after learning (the most kurtotic and the least kurtotic sources only).

Implementation Notes: Important!

- The implementation of the rigid-body learning equations should meet some requirements:
  - 1) easy representation of the involved matrix-quantities,
  - 2) spare use of arrays and computations,
  - 3) accurate integration of the differential equations.

- Previous solutions:
  - 1) requirement inherently met,
  - 2) arrays used as they are: requirement not met,
  - 3) Euler method for matrix **B**(t): allowed because it belongs to a linear space. Exponential map for matrix **W**(t) updating:

$$\mathbf{W}((n+1)h) = \exp(h\mathbf{B}(n))\,\mathbf{W}(nh)$$

computationally expensive if used as is (2 not met), inaccurate if approximated by the Taylor series: requirement only partially met.

Stressing out the Importance of Integration

- Implementing a differential learning equation means discretizing it in time.

- The discretization process in not consequence-free:
  - The properties of the continuous-time equations (such as intrinsic stability) generally do not preserve.
  - The invariants generally break.

- Main point: The integration process – which actually is what we call "algorithm" – should be definitely paid attention to !!

  - Some numerical examples in: S. Fiori, "A Minor Subspace Algorithm Based on Neural Stiefel Dynamics", Int. J. of Neural Systems, Vol. 12, No. 5, pp. 339 – 350, 2002

First: Reworking the equations

- Problem ① : The state-matrix **B**(t) is always of maximum dimension $p \times p$ even when $m << p$.
- Solution: Reformulate the equations as:

$$\begin{cases} \dfrac{d\mathbf{W}}{dt} = \mathbf{V} , \\ \dfrac{d\mathbf{V}}{dt} = g(\mathbf{V}, \mathbf{W}). \end{cases}$$

- In other terms, instead of using the Lie-algebra parameterization, the tangent-space parameterization is chosen for the equations.
- Advantage: The new parameterization is computationally cheaper (**V** has dimension $p \times m$).

Main result on Reworking the Equations

- Any element **V** tangent to the Stiefel manifold at **W** can be written in the form (Celledoni and Owren, 2001):

$$\mathbf{V} = (\mathbf{G}\mathbf{W}^T - \mathbf{W}\mathbf{G}^T)\mathbf{W}$$

- By using the previous *ansatz*:

$$g(\mathbf{V},\mathbf{W}) = \mathbf{F} + \mathbf{P} + \mathbf{W}((\mathbf{F}+\mathbf{P})^T\mathbf{W}) + \mathbf{G}(\mathbf{W}^T\mathbf{V}) - \mathbf{W}(\mathbf{G}^T\mathbf{V})$$

- where $\mathbf{P} = -n\mathbf{V}$ and $\mathbf{G} = \mathbf{V} - \mathbf{W}(\mathbf{W}^T\mathbf{V})/2$.
- Note: The parentheses in the above equations suggest a computationally-effective way to calculate the matrix-products.

- Reference: E. Celledoni and B. Owren, On the implementation of Lie-group methods on the Stiefel manifold, Preprint Numerics no. 9/2001, Norwegian University of Science and Technology, Trondheim (Norway), 2001

Second: Integrate the Equations

- Problem ② : The differential equations should be integrated properly.
- Solution under investigation: Use Geometric Integration (GI) methods.
- In the present case, the differential equation for **V** may be solved by the standard Euler method, because it belongs to the tangent space to the Stiefel manifold at **W**, which is a linear space.
- We integrate the differential equation for **W** using the Lie-Euler method which advances the numerical solution by using the left transitive action of SO(p) on the Stiefel manifold. The action is lifted to the Lie algebra **so**(p) using the exponential map.

BTW: What is Geometric Integration ? (1)

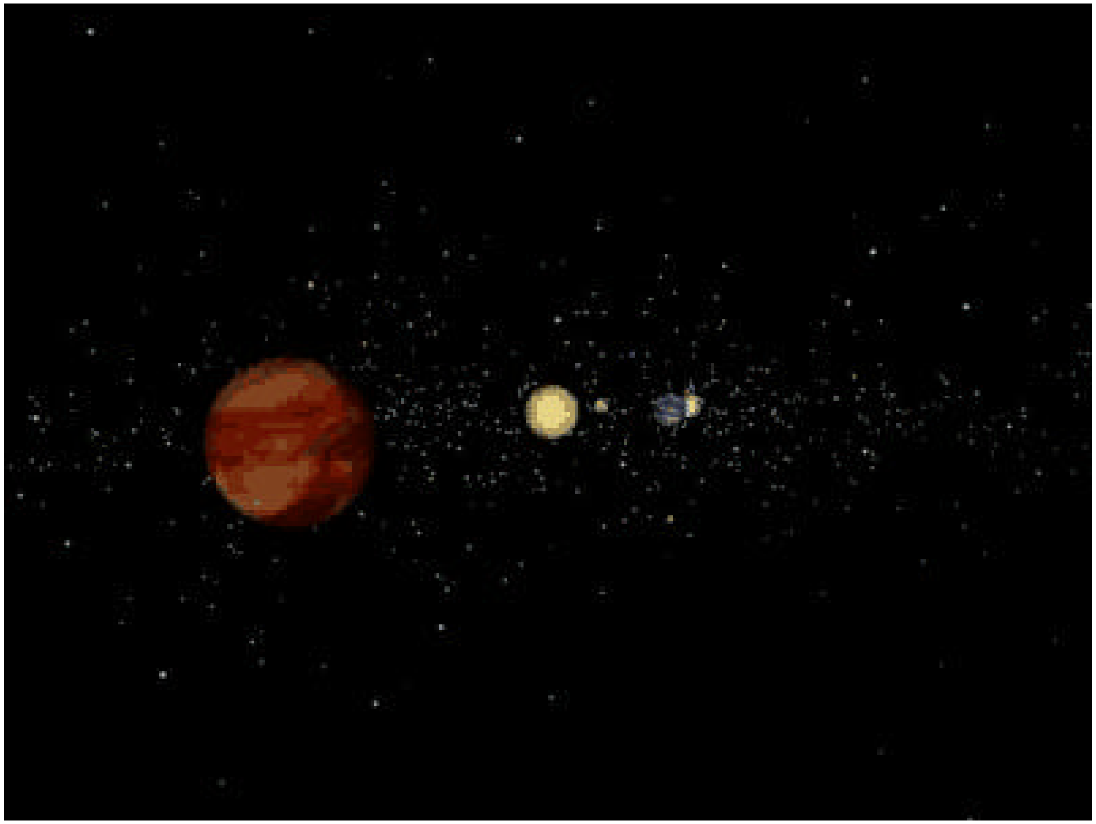
- Geometric integration is a new approach to simulating the motion of large systems.
- The new methods are faster, more reliable, and often simpler than traditional approaches.
- They are being used in the structure of liquids, polymers and bio-molecules, quantum mechanics and nano-devices, biological models, chemical reaction-diffusion systems, the dynamics of flexible structures, and several more.

What is Geometric Integration ? (2)

- Although diverse, the above systems preserve some underlying geometric structure which influences the qualitative nature of the phenomena they produce. In geometric integration these properties are built into the numerical method, which gives the method markedly superior performance, especially during long simulations.
- Main references:
  - E. Hairer, C. Lubich, and G. Wanner, Geometric Numerical Integration, *Springer series in Computational Mathematics*, Springer, 2002
  - P.J. Olver, Applications of Lie Groups to Differential Equations, *Springer series in Graduate Text in Mathematics*, Springer, 1993 (Second Edition)
  - Geometric integration interest group web page: http://www.focm.net/gi/

An Illustrative Example

- **Full many-particle simulation of Near Earth Asteroids**



- **Source: *Sverker Edvardsson* (http://www.fmi.mh.se/~sverkere/)**

New Learning Algorithm

- The new learning algorithm (under construction) writes:

$$\begin{cases} \mathbf{V}_{n+1} = \mathbf{V}_n + hg(\mathbf{V}_n, \mathbf{W}_n), \\ \mathbf{G}_n = \mathbf{V}_n - \mathbf{W}_n(\mathbf{W}_n^T \mathbf{V}_n)/2, \\ \mathbf{W}_{n+1} = \exp(h(\mathbf{G}_n \mathbf{W}_n^T - \mathbf{W}_n \mathbf{G}_n^T))\mathbf{W}_n. \end{cases}$$

- The exponentiation is much expensive in terms of computation cost. Some alternate methods as well as numerical tricks appeared in:
  - E. Celledoni and A. Iserles, Approximating the exponential form of a Lie algebra to a Lie group, Math. Comp. 69, pp. 1457 - 1480, 2000
  - E. Celledoni and S. Fiori, Neural learning by geometric integration of reduced ‘rigid-body’ equations. Preprint Numerics no. 4/2002, Norwegian University of Science and Technology, Trondheim (Norway), 2002
  - I. Yamada and T. Ezaki, An orthogonal matrix optimization by dual Cayley parameterization technique. To appear on ICA*2003 Proceedings.

Conclusions - 1

- Neural learning with orthonormality constraints has been surveyed.
- The formulation of the learning differential equations and of the invariants is of prime importance.
- The correct choice of the most proper integration algorithm looks truly an important part of learning algorithm design.

Conclusions - 2

- It seems worth to consider a unifying view among different research streams aiming at preserving invariants, namely:
  - 1) Geodesic learning (Edelman, Arias and Smith, 1998; Nishimori 1999).
  - 2) Projection methods (Manton, 2001)
  - 3) Lagrange multiplier methods for gradient adaptation (for a discussion see Douglas, Amari and Kung, 1999)
  - 4) Fixed-point iteration method (? – It stems from 3)

Further reading...

- A. Cichocki and P. Georgev, *Blind separation algorithms with matrix constraints*, IEICE Trans. on Fundamentals. To appear
- S.C. Douglas, S.-I. Amari and S.-Y. Kung, *Gradient adaptation with unit-norm constraints*, Technical report of the Department of Electrical Engineering, School of Engineering and Applied Science, Southern Methodist University, Dallas (TX, USA). February 1999
- A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett and A. Zanna: Lie-group methods, Acta Numerica, Vol. 9, pp. 215 - 365, 2000
- J.H. Manton, *Optimisation algorithms exploiting unitary constraints*, IEEE Transactions on Signal Processing. To appear
- T. Rapcsák, *On minimization on Stiefel manifold*, European Journal of Operational Research. To appear

Acknowledgments and Thanks...

Even if it is impossible to acknowledge adequately the help of the following persons, I would like to thank sincerely:

Prof. S.-I. Amari for the invitation to give the lecture and for his unfailing example.

Dr. Y. Nishimori and his colleagues for the invitation to visit the AIST institute and for making the stay definitely pleasant.

Mrs. Nishimori for the kind hospitality and for the wonderful curry rice.