Adjoint Orbits, Principal Components, and Neural Nets

- Some facts about Lie groups and examples
- Examples of adjoint orbits and a distance measure
- Descent equations on adjoint orbits
- Properties of the double bracket equation
- Smoothed versions of the double bracket equation
- The principal component extractor
- The performance of subspace filters
- Variations on a theme

Where We Are

9:30 - 10:45	Part 1. Examples and Mathematical Background
10:45 - 11:15	Coffee break
11:15- 12:30	Part 2. Principal components, Neural Nets, and Automata
12:30 - 14:30	Lunch
14:30 - 15:45	Part 3. Precise and Approximate Representation of Numbers
15:45 - 16:15	Coffee break
16:15 - 17:30	Part 4. Quantum Computation

The Adjoint Orbit Theory and Some Applications

1. Some facts about Lie groups and examples
- Examples of adjoint orbits and a distance measure
- Descent equations on adjoint orbits
- Properties of the double bracket equation
- Smoothed versions of the double bracket equation
- Loops and deck transformations

Some Background

By a Lie Group G we understand a group with a topology such that multiplication and inversion are continuous. (In this setting continuous implies differentiable.)

We say that a group acts on a differentiable manifold X via ϕ if $\phi : G \times X \to M$ is differentiable and $\phi(G_2 G_1, x) = \phi(G_2, \phi(G_1, x))$.

The group of orthogonal matrices $So(n)$ acts on the $n-1$-dimensional sphere via the action $\phi(\Theta, x) = \Theta x$

More Mathematics Background

Associated with every Lie group is a Lie algebra L which may be thought of as describing how G looks in a small set around the identity. Abstractly, a Lie algebra is a vector space with a bilinear mapping $\phi : L \times L \mapsto L$ such that

$$[L_1, L_2] = -[L_2, L_1]$$

$$[L_1, [L_2, L_3]] + [L_2, [L_3, L_1]] + [L_3, [L_1, L_2]] = 0$$

The Lie algebra associated with the real orthogonal group is the set of skew- symmetric matrices of the same dimension. The bilinear operation is given by $[\Omega_1, \Omega_2] = \Omega_1\Omega_2 - \Omega_2\Omega_1$.

A Little More Mathematics Background

Let Θ be an orthogonal matrix and let Q be a symmetric matrix with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$. The formula $\Theta^T Q \Theta$ defines a group action on $\mathrm{Sym}(\lambda_1, \lambda_2, ..., \lambda_n)$. The set of orthogonal matrices is of dimension $n(n-1)/2$ and the space $\mathrm{Sym}(\Lambda)$ is of dimension $n(n+1)/2$. This action is basic to a lot Matlab!

The action of the group of unitary matrices on the space of skew-hermitian matrices via $(U, H) \mapsto U^\dagger H U$ can be thought of as generalizing this action. It is an example of a group acting on its own Lie algebra. This is an adjoint action.

Still More Mathematics Background

Consider Lie algebras whose elements are n by n matrices and Lie groups whose elements are nonsingular n by n matrices. The mapping exp : $L \mapsto e^L$ sends the Lie algebra into the group of invertible matrices. The identity $P^{-1}e^L P = e^{P^{-1}LP}$ defines the adjoint action.

If $\phi : G \times X \to X$ is a group action then there is an equivalence relation on X defined by $x \approx y$ if $y = \phi(G_1, x)$ for some $G_1 \in G$. Sets of equivalent points are called orbits. The subset of $H \subset G$ such that $\phi(H, x_0) = x_0$ forms a subgroup called the isotropy group t x_0.

The Last for now, Mathematics Background

Any $L_1 \in L$ defines via $[L_1, \cdot] : L \to L$, a linear transformation on a finite dimensional space. It is often written $\mathrm{ad}_{L_1}(\cdot)$. $\mathrm{ad}_{L_1}(\mathrm{ad}_{L_2}(\cdot))$ $= [L_1, [L_2, \cdot]]$ defines a linear transformation on L as well. The sum of the eigenvalues of this map defines what is called the Killing form $\kappa(L_1, L_2)$, on L. For semisimple compact groups such as the orthogonal or special unitary group, the Killing form is negative definite and proportional to the more familiar $\mathrm{tr}(\Omega_1 \Omega_2)$.

The Killing form on G defines a metric on the adjoint orbit called the normal metric.

Getting a Feel for the Normal Metric

Explanation: Consider perturbing Θ via $\Theta \mapsto \Theta(I + \Omega)$. Linearizing the equation

$$\Theta^T Q \Theta = H$$

we get

$$H\Omega + \Omega^T H = [H, \Omega] = dH$$

Thus

$$\Omega = \mathrm{ad}_H^{-1}(dH)$$

If H is diagonal then

$$\omega_{ij} = \frac{dh_{ij}}{\lambda_i - \lambda_j}$$

Steepest Descent on an Adjoint Orbit

Let $Q = Q^T$ and $N = N^T$ be symmetric matrices and let Θ be orthogonal. Consider the function $\mathrm{tr}\Theta^T Q \Theta N$ thought of as a function on the orthogonal matrices. Relative to the Killing metric on the orthogonal group, the gradient descent flow for minimizing this function is

$$\dot{\Theta} = [\Theta^T Q \Theta, N]\Theta$$

If we let $\Theta^T Q \Theta = H$ then the derivative of H can be expressed as

$$\dot{H} = [H, [H, N]]$$

A Descent Equation on an Adjoint Orbit

Let $Q = Q^T$ and $N = N^T$ be symmetric matrices and let $\psi(H)$ be a real valued function on $\mathrm{Sym}(\Lambda)$. What is the gradient of $\psi(H)$? The gradient on a Riemannian space is $G^{-1}d\psi$. On $\mathrm{Sym}(\Lambda)$ the inverse of the Riemannian metric is given by $[H,[H,\cdot]]$. and so the descent equation is

$$\dot{H} = -[H,[H,d\psi(H)]]$$

Thus for $\psi(H) = \mathrm{tr}(HN)$ we have $\dot{H} = -[H,[H,N]]$. If N is diagonal then $\mathrm{tr}HN$ achieves its minimum when H is diagonal and similarly ordered with $-N$.

A Descent Equation with Multiple Equilibria

If $\psi(H) = -\mathrm{tr}(\mathrm{diag}(H)H)$ then $\dot{H} = [H,[H,2\mathrm{diag}(H)]]$ Let $Q = Q^T$ and $N = N^T$ be diagonal matrices with distinct eigenvalues. The descent equation is

$$\dot{H} = -[H,[H,d\psi(H)]]$$

Thus for $\psi(H) = \mathrm{tr}(HN)$ we have $\dot{H} = [H,[H,N]]$
If $\psi(H) = \mathrm{diag}H$ then $\dot{H} = 2[H,[H,\mathrm{diag}(H)]]$

A Descent Equation with Smoothing Added

Consider replacing the system

$$\dot{H} = [H, [H, N]]$$

with

$$\dot{H} = [H, q(D)P] \quad ; \quad p(D)P = [H, N]$$

Here $D = d/dt$. This smooths the signals but does not alter the equilibrium points. Stability is un affected if q/p is a positive real function.

The Double Bracket Flow for Analog Computation

Principal Components in R^n

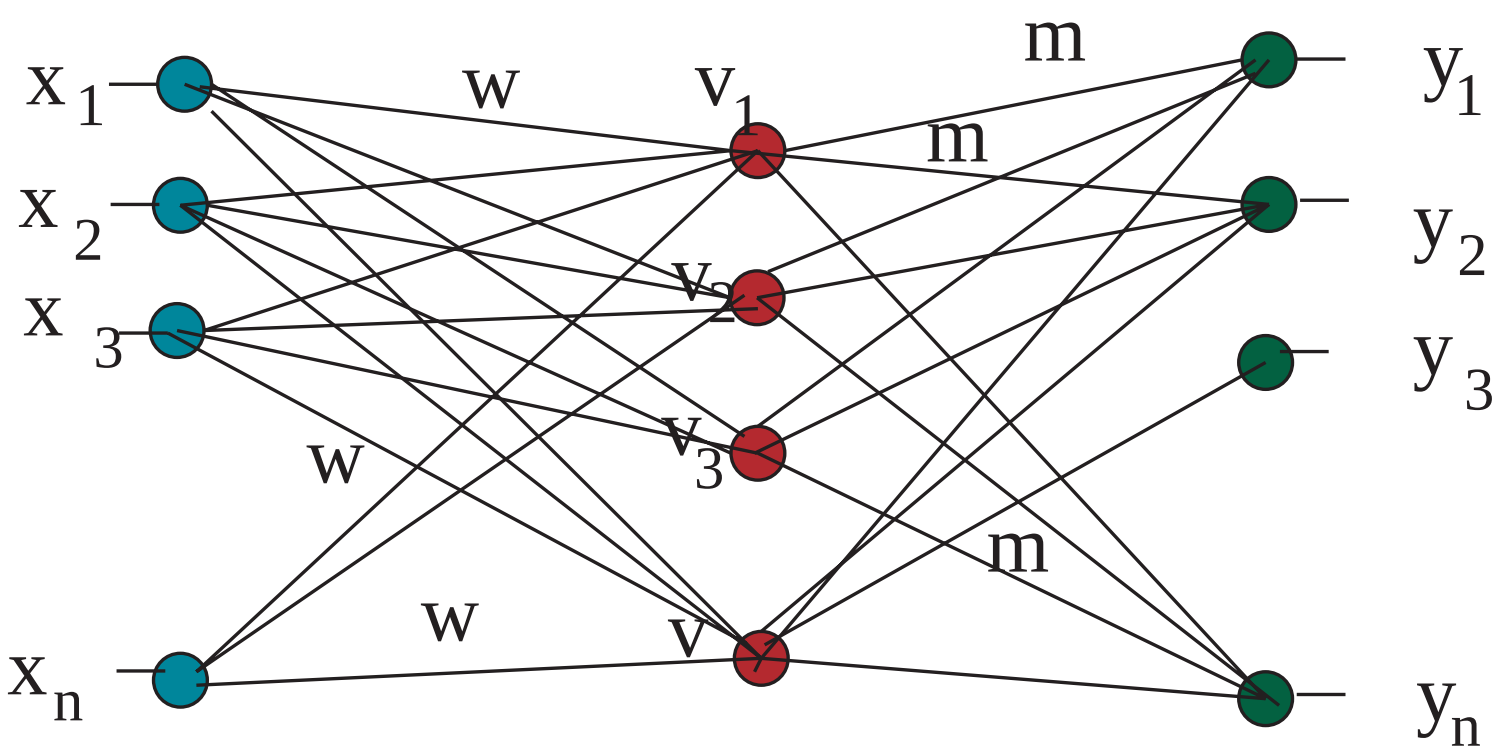
Learning without a teacher is sometimes approached by finding principal components.

$\dot{W} = x(t)x^T(t)$ - forgetting term

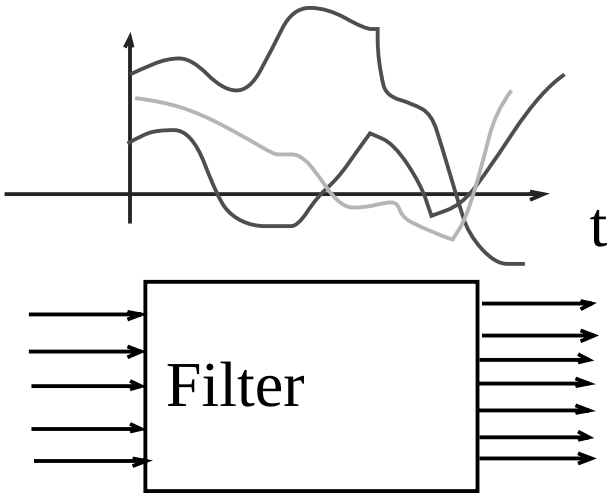
$\Theta^T(t)W(t)\Theta(t) = \text{diag}(\lambda_1 ..., \lambda_n)$

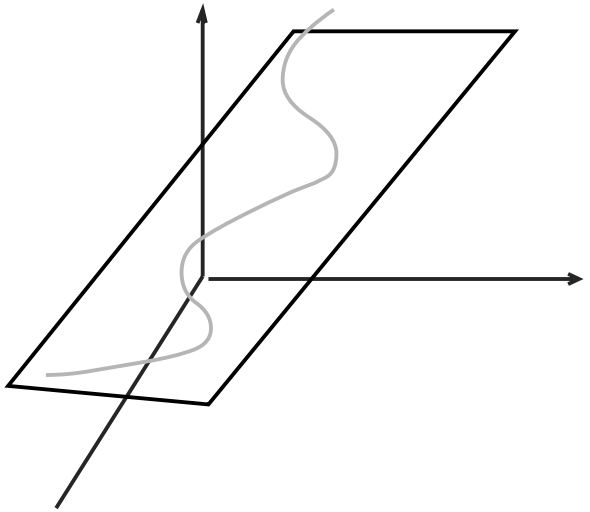
Columns of Θ are "components'

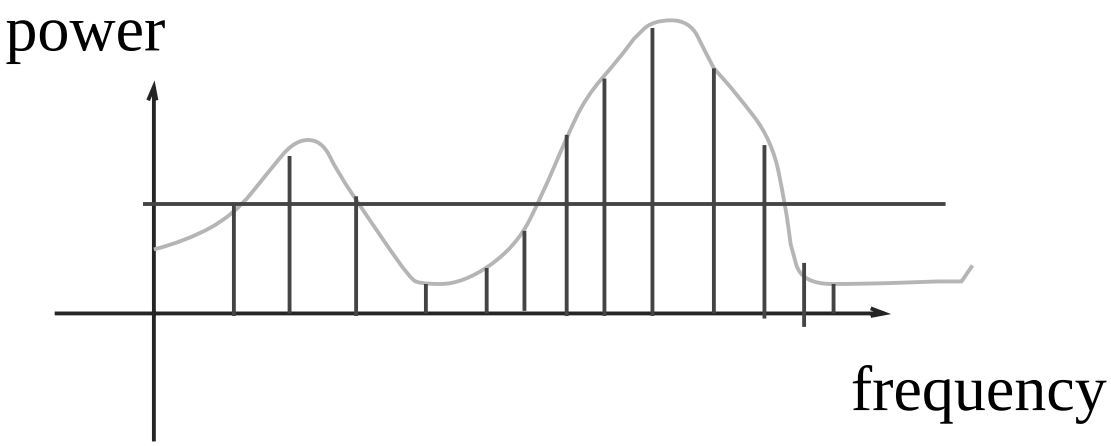
The principal components are assembled in a hidden layer


x 1
x 2
x 3
x n
w
w
w
v 1
v 2
v 3
v
m
m
m
y 1
y 2
y 3
y n

Adaptive Subspace Filtering







Some Equations

Let u be a vector of inputs, and let Λ be a diagonal "editing" matrix that selects energy levels that are desirable. An adaptive subspace filter with input u and output y can be realized by implementing the equations

$$\frac{dQ}{dt} = uu^T - (1 - tr(Q))Q$$

$$\frac{d\Theta}{dt} = [\Theta Q \Theta^T, N]\Theta$$

$$y = \Theta \Lambda \Theta^T u$$

Neural Nets as Flows on Grassmann Manifolds

Denote by $G(n,k)$ the space of k-planes in n-space. This space is a differentiable manifold that can be parameterized by the set of all k by n matrices of rank k. It is a manifold. Adaptive subspace filters steer the weights so as to define a particular element of this space. Thus $\Lambda\Theta$, defines such a point if Λ looks like

$$\Lambda = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{bmatrix}$$

Summary of Part 2

1. We have given some mathematical background necessary to work with flows on adjoint orbits and indicated some applications.

2. We have defined flows that will stabilize at invariant subspaces corresponding to the principal components of a vector process. These flows can be interpreted as flows that learn without a teacher.

3. We have argued that in spite of its limitations, steepest descent is usually the first choice in algorithm design.

4. We have interpreted a basic neural network algorithm as a flow in a Grassmann manifold generated by a steepest descent tracking algorithm.

A Few References

M. W. Berry et al., “Matrices, Vector Spaces, and Information Retrieval” SIAM Review, vol. 41, No. 2, 1999.

R. W. Brockett, “Dynamical Systems That Learn Subspaces” in Mathematical System Theory: The Influence of R. E. Kalman, (A.C. Antoulas, ed.) Springer -Verlag, Berlin. 1991. pp. 579--592.

R. W. Brockett “An Estimation Theoretic Basis for the Design of Sorting and Classification Networks,” in Neural Networks, (R. Mammone and Y. Zeevi, eds.) Academic Press, 1991, pp. 23-41.