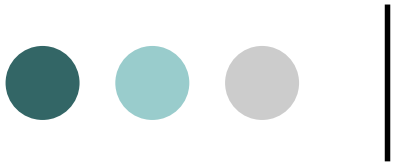


Applications of geometric optimisation techniques to engineering problems

Jochen Trumpf

Jochen.Trumpf@anu.edu.au

Department of Information Engineering

Research School of Information Sciences and Engineering

The Australian National University

and

National ICT Australia Ltd.

overview

- **What is geometric optimisation?**

overview

- **What is geometric optimisation?**
- **Ex 1: Blind Source Separation (BSS)**

overview

- **What is geometric optimisation?**
- **Ex 1: Blind Source Separation (BSS)**
- **Independent Component Analysis (ICA)**

overview

- What is geometric optimisation?
- Ex 1: Blind Source Separation (BSS)
- Independent Component Analysis (ICA)
- Ex 2: face recognition

overview

- What is geometric optimisation?
- Ex 1: Blind Source Separation (BSS)
- Independent Component Analysis (ICA)
- Ex 2: face recognition
- dominant eigenspaces of matrix pencils (LDA)

overview

THE AUSTRALIAN NATIONAL UNIVERSITY

- What is geometric optimisation?
- Ex 1: Blind Source Separation (BSS)
- Independent Component Analysis (ICA)
- Ex 2: face recognition
- dominant eigenspaces of matrix pencils (LDA)
- Ex 3: time series clustering

NATIONAL ICT AUSTRALIA

overview

THE AUSTRALIAN NATIONAL UNIVERSITY

- What is geometric optimisation?
- Ex 1: Blind Source Separation (BSS)
- Independent Component Analysis (ICA)
- Ex 2: face recognition
- dominant eigenspaces of matrix pencils (LDA)
- Ex 3: time series clustering
- “on-the-fly” geometry

NATIONAL ICT AUSTRALIA

overview

THE AUSTRALIAN NATIONAL UNIVERSITY

- What is geometric optimisation?
- Ex 1: Blind Source Separation (BSS)
- Independent Component Analysis (ICA)
- Ex 2: face recognition
- dominant eigenspaces of matrix pencils (LDA)
- Ex 3: time series clustering
- “on-the-fly” geometry
- state of the art and open problems

NATIONAL ICT AUSTRALIA

What is geometric optimisation?

THE AUSTRALIAN NATIONAL UNIVERSITY

Given a real valued function

$$f : M \longrightarrow \mathbb{R},\ x \mapsto f(x)$$

defined on some geometric object M, here a smooth manifold, find a method to compute (if it exists)

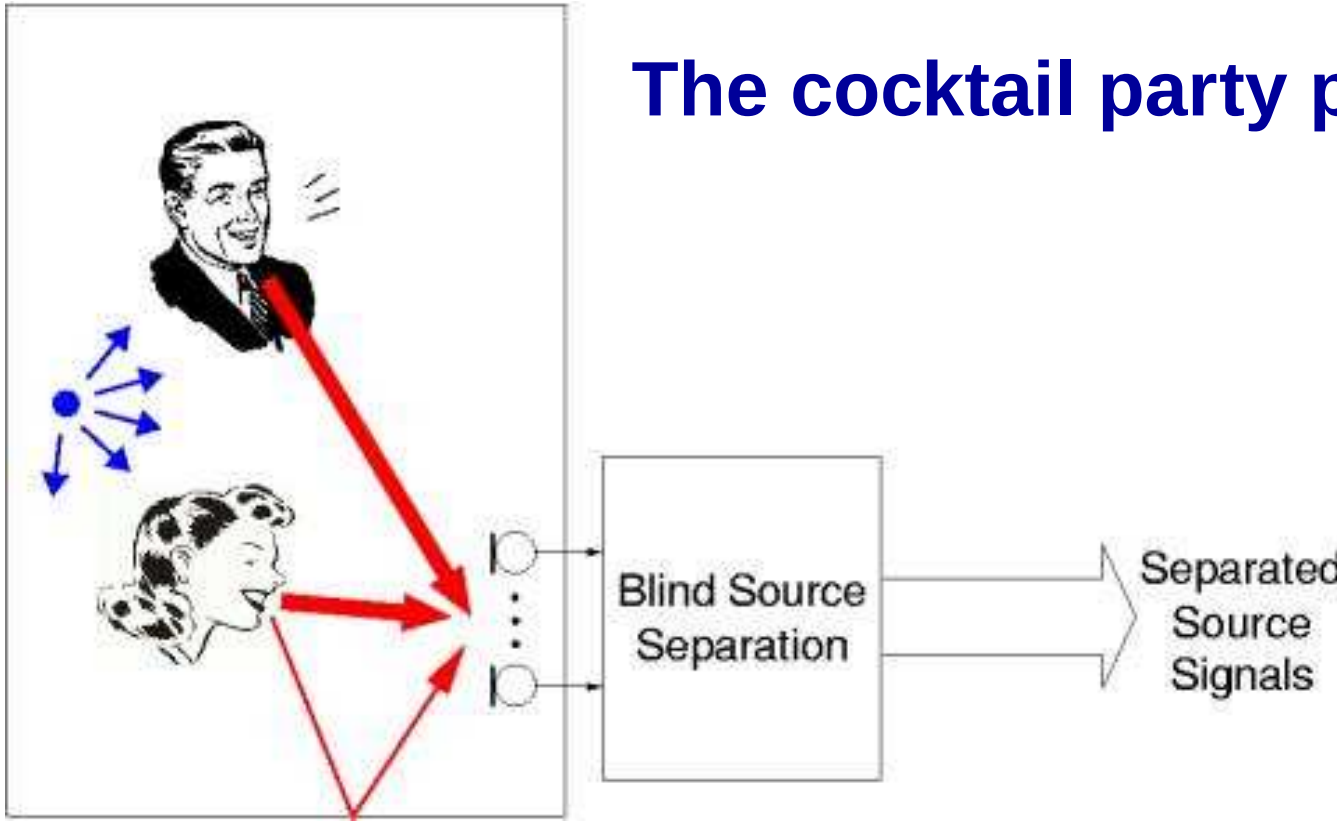
$$x_* := \operatorname*{argmin}_{x \in M} f(x)$$

that utilises the (local) geometry of M.

NATIONAL ICT AUSTRALIA

Ex 1: Blind Source Separation

THE AUSTRALIAN NATIONAL UNIVERSITY

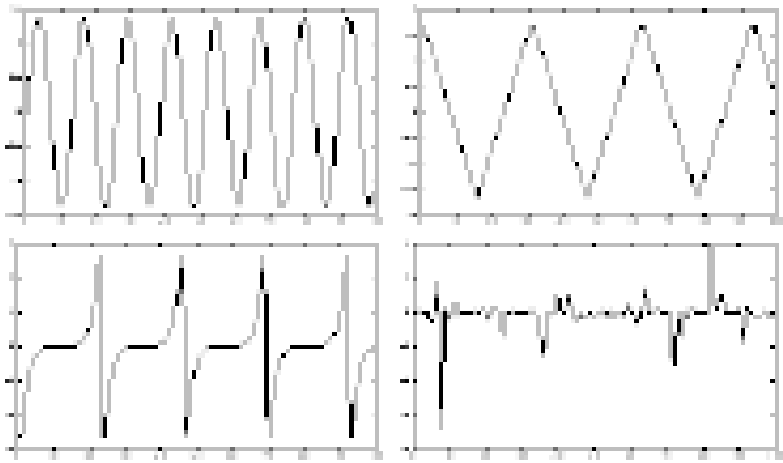


The cocktail party problem.

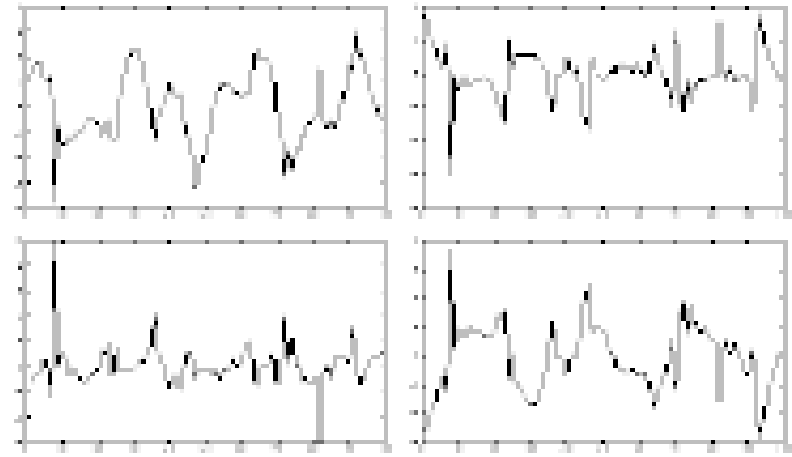
NATIONAL ICT AUSTRALIA

Image: http://www.lnt.de/LMS/research/projects/BSS

Ex 1: Blind Source Separation



source signals



observed mixtures

audio, EEG, MEG, fMRI, wireless, ...

Image: http://www.cis.hut.fi/aapo/papers/NCS99web/node17.html

BSS – the model

Individual signals ($i = 1, \ldots, d$)

$$x_i : [0, T] \longrightarrow \mathbb{R},\ t \mapsto x_i(t)$$

are being uniformly sampled and the samples collected into row vectors

$$x_i = \begin{pmatrix} x_i(t_0) & x_i(t_0 + \Delta) & \ldots & x_i(t_0 + (N-1) \cdot \Delta) \end{pmatrix}$$

which are then stacked into a matrix

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d \times N}.$$

BSS – the model

It is assumed that there are as many source signals as observed signals and that they are related by

$$X_o = M \cdot X_s$$

where $X_o, X_s \in \mathbb{R}^{d \times N}$ **and** $M \in \mathrm{GL}_d(\mathbb{R})$**.**

BSS – the model

It is assumed that there are as many source signals as observed signals and that they are related by

$$X_o = M \cdot X_s$$

where $X_o, X_s \in \mathbb{R}^{d \times N}$ **and** $M \in \mathrm{GL}_d(\mathbb{R})$**.**

Task: Find X_s **(or** M^{-1}**) from knowing** X_o **subject to some plausible criterion.**

BSS as ICA problem

We treat the columns of X_o as i.i.d. samples of an observed random variable vector Y given by

$$Y = M \cdot X$$

where X is the unknown random variable source vector.

BSS as ICA problem

THE AUSTRALIAN NATIONAL UNIVERSITY

We treat the columns of X_o as i.i.d. samples of an observed random variable vector Y given by

$$Y = M \cdot X$$

where X is the unknown random variable source vector.

The ICA paradigm is now that the components of X, i.e. the individual signals, are mutually independent.

NATIONAL ICT AUSTRALIA

BSS as ICA problem

Hence, we are trying to find the invertible M that makes the components of the corresponding X "as independent as possible".

BSS as ICA problem

Hence, we are trying to find the invertible M that makes the components of the corresponding X "as independent as possible".

Note: The matrix M in

$$Y = M \cdot X$$

is identifiable up to scaling and permutations if and only if the components of X are mutually independent and at most one of them is Gaussian.

BSS as ICA problem

A computational trick is centering and prewhitening: multiply by the square root of the covariance matrix of Y (assuming finite second moments) to obtain

$$Y = Q \cdot X$$

where $Q \in \mathrm{O}_d(\mathbb{R})$ and X and Y are zero mean and unit variance.

BSS as ICA problem

A computational trick is centering and prewhitening: multiply by the square root of the covariance matrix of Y (assuming finite second moments) to obtain

$$Y = Q \cdot X$$

where $Q \in \mathrm{O}_d(\mathbb{R})$ and X and Y are zero mean and unit variance.

Note: Prewhitening from samples works best in the Gaussian case ...

see IEEE TSP, 53(10):3625–3632, 2005

ICA as geometric optimisation problem

We arrive at the geometric optimisation problem of minimising mutual information between the components of $Q^\top Y$ over $Q \in \mathrm{O}_d(\mathbb{R})$.

ICA as geometric optimisation problem

We arrive at the geometric optimisation problem of minimising mutual information between the components of $Q^\top Y$ over $Q \in \mathrm{O}_d(\mathbb{R})$.

One-unit FastICA maximises $\mathbb{E}[G(q^\top Y)]$ over $q \in S^{d-1}$ where $G : \mathbb{R} \longrightarrow \mathbb{R}$, $z \mapsto \frac{1}{a} \log \cosh(az)$ is a contrast function.

The expectation is computed from samples, the optimisation method is an approximate Newton on manifold algorithm.

http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb

Ex 2: face recognition

Image: IEEE TPAMI, 23(2):228–233, 2001

face recognition – the model

An image is represented as a vector $X \in \mathbb{R}^t$. Images are divided in c classes with N_j images X_i^j, $i = 1, \ldots, N_j$ in class $j = 1, \ldots, c$.

face recognition – the model

An image is represented as a vector $X \in \mathbb{R}^t$. Images are divided in c classes with N_j images X_i^j, $i = 1, \ldots, N_j$ in class $j = 1, \ldots, c$.

Consider the within-class scatter matrix

$$S_w = \sum_{i,j} (X_i^j - \mu_j)(X_i^j - \mu_j)^\top$$

and the between-class scatter matrix

$$S_b = \sum_{j} (\mu_j - \mu)(\mu_j - \mu)^\top.$$

face recognition as LDA problem

THE AUSTRALIAN NATIONAL UNIVERSITY

Orthogonally projecting the image vectors into a lower dimensional space $Y = Q^\top X$ yields projected scatter matrices $Q^\top S_{\{w,b\}} Q$.

NATIONAL ICT AUSTRALIA

face recognition as LDA problem

THE AUSTRALIAN NATIONAL UNIVERSITY

Orthogonally projecting the image vectors into a lower dimensional space $Y = Q^\top X$ yields projected scatter matrices $Q^\top S_{\{w,b\}} Q$.

The aim is to maximise $\frac{\det(Q^\top S_b Q)}{\det(Q^\top S_w Q)}$ over $Q \in \mathrm{St}(d,t)$, the orthogonal Stiefel manifold.

NATIONAL ICT AUSTRALIA

face recognition as LDA problem

THE AUSTRALIAN NATIONAL UNIVERSITY

Orthogonally projecting the image vectors into a lower dimensional space $Y = Q^\top X$ yields projected scatter matrices $Q^\top S_{\{w,b\}} Q$.

The aim is to maximise $\frac{\det(Q^\top S_b Q)}{\det(Q^\top S_w Q)}$ over $Q \in \mathrm{St}(d,t)$, the orthogonal Stiefel manifold.

This amounts to finding the dominant d-dimensional eigenspace of the pencil (S_b, S_w).

NATIONAL ICT AUSTRALIA

LDA as geometric optimisation problem

Given a symmetric/positive-definite matrix pencil (A, B) with eigenvalues ($Ax = \lambda Bx$) $\lambda_1 \geq \cdots \geq \lambda_d > \lambda_{d+1} \geq \cdots \geq \lambda_n$ the unique d-dimensional dominant eigenspace is the unique global maximum of

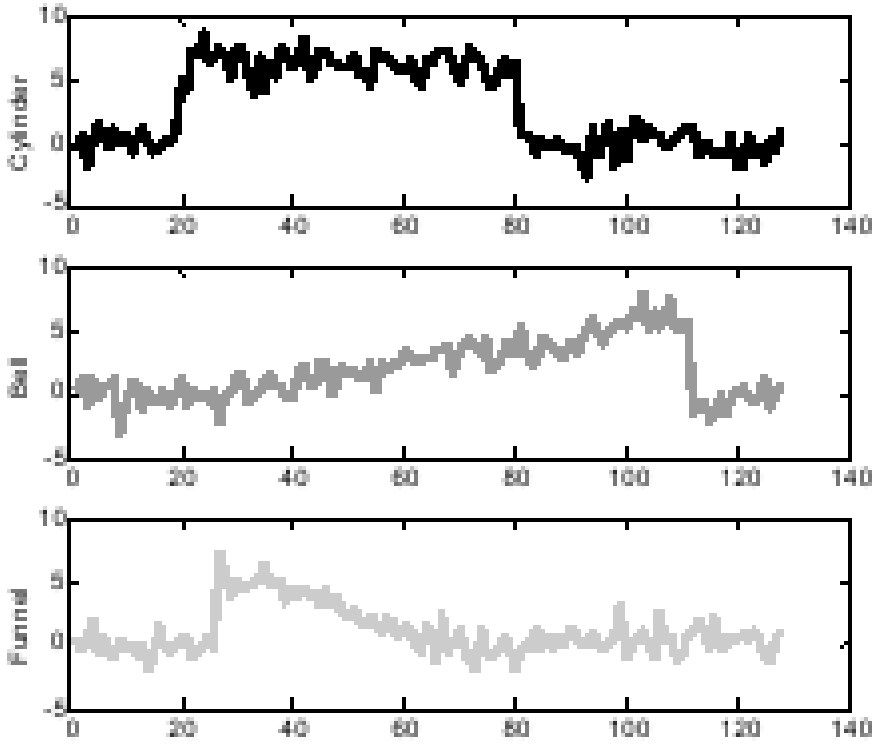
$$f : \mathrm{Grass}(d, n) \longrightarrow \mathbb{R}, \ [Q] \mapsto \mathrm{tr}(Q^\top AQ(Q^T BQ)^{-1})$$

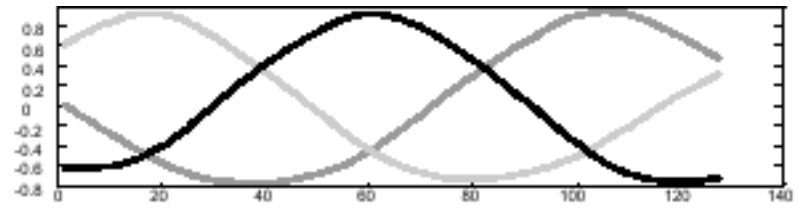
Ex 3: time-series clustering

A time series is a (finite) sequence $\{x_t\}_{t=1,\ldots,N}$ of vectors (in $\mathbb{R}^n$), e.g. arising from (sampling) a trajectory of a dynamical system.
A popular method of time-series clustering works in delay space

$$\left\{ \left(\begin{array}{c} x_p \\ x_{p-1} \\ \vdots \\ x_{p-l+1} \end{array} \right) \middle| \; p = l, \ldots, N \right\}$$

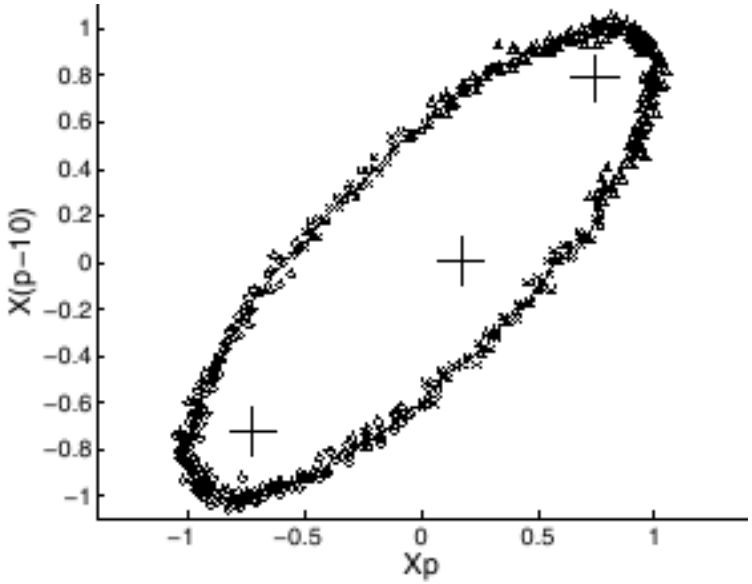
Ex 3: time-series clustering

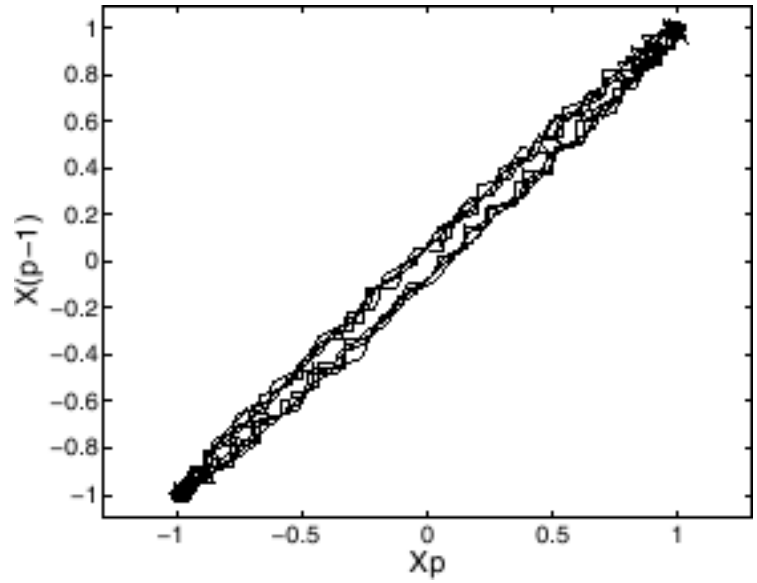




Knowl. Inf. Syst., 8(2):154-177, 2005

Ex 3: time-series clustering





ICDM 2005, pp. 114–121

state of the art

Let $\mathcal{M}$ be a d-dimensional Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}$ be smooth.

The derivative of f at $x \in \mathcal{M}$ is a linear form

$$\mathrm{D}\, f(x) : T_x\mathcal{M} \to \mathbb{R}$$

A point $x^* \in \mathcal{M}$ is called a critical point of f if

$$\mathrm{D}\, f(x^*)\xi = 0, \qquad \forall \xi \in T_{x^*}\mathcal{M}.$$

state of the art

Fact: $x^* \in \mathcal{M}$ is a strict local minimum of f if

(a) x^* is a critical point of f,

(b) the Hessian form

$$\operatorname{hess} f(x^*) : T_{x^*}\mathcal{M} \times T_{x^*}\mathcal{M} \to \mathbb{R}$$

is positive definite.

state of the art

Fact: $x^* \in \mathcal{M}$ **is a strict local minimum of** f **if**

(a) x^* **is a critical point of** f**,**

(b) the Hessian form

$$\operatorname{hess} f(x^*) : T_{x^*}\mathcal{M} \times T_{x^*}\mathcal{M} \to \mathbb{R}$$

is positive definite.

Geodesics of $\mathcal{M}$**:** $\forall x \in \mathcal{M}$ **and** $\xi \in T_x\mathcal{M}$

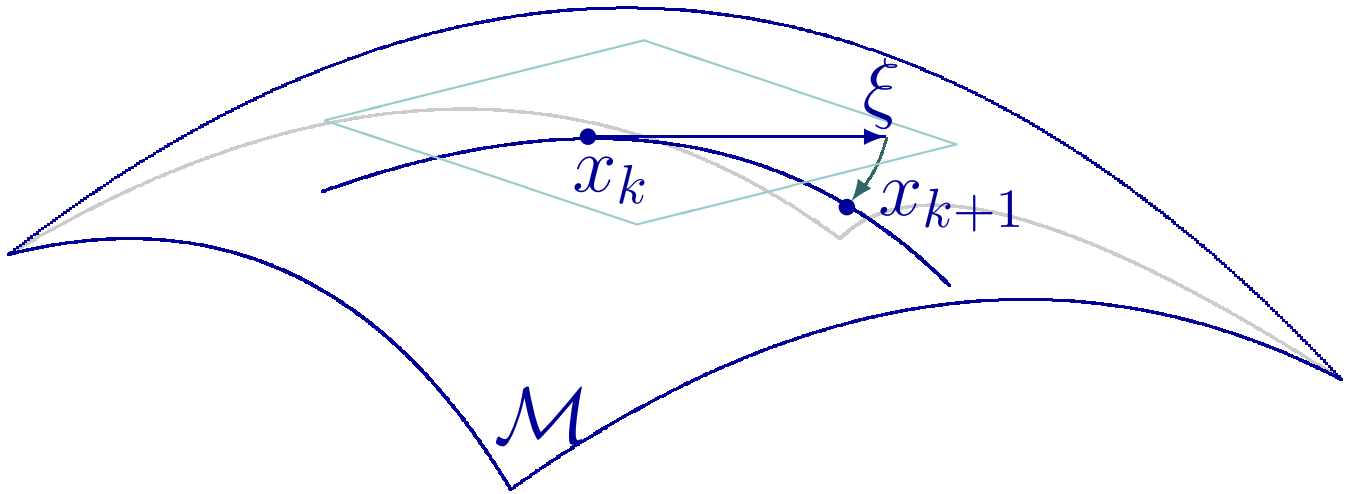
$$\gamma_x : \mathbb{R} \supset (-\varepsilon, \varepsilon) \to \mathcal{M}, \qquad \varepsilon \mapsto \gamma_x(\varepsilon)$$

such that $\gamma_x(0) = x$ **and** $\dot{\gamma}_x(0) = \xi$**.**

state of the art

Riemannian Newton direction $\xi \in T_x\mathcal{M}$ by solving

$$\operatorname{hess} f(x) \cdot \xi = \operatorname{grad} f(x)$$



state of the art

Local parameterisation of $\mathcal{M}$ around $x \in \mathcal{M}$

$$\mu_x : \mathbb{R}^d \to \mathcal{M}, \quad \kappa \mapsto \mu_x(\kappa); \qquad \mu_x(0) = x$$

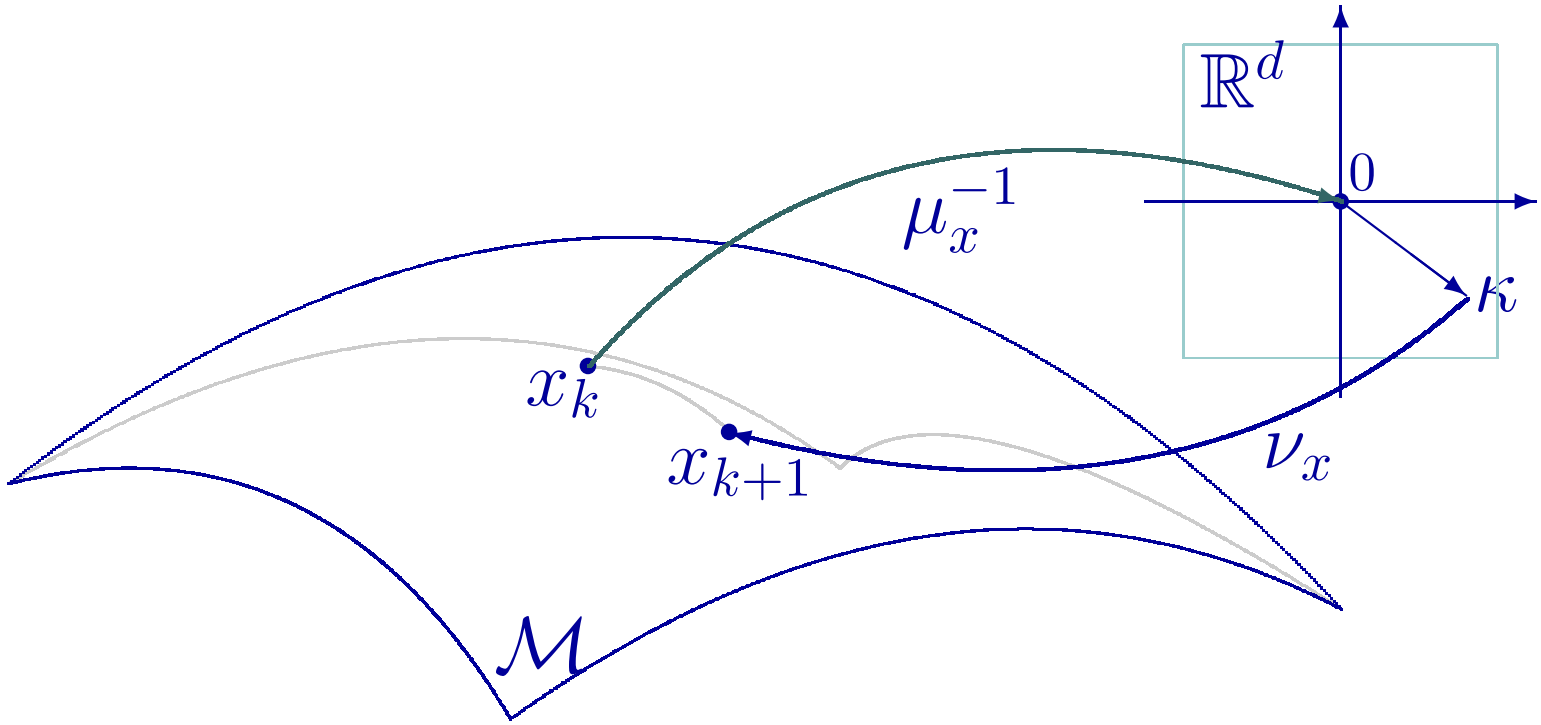
Construct locally

$$f \circ \mu_x : \mathbb{R}^d \to \mathbb{R}$$

Euclidean Newton direction $\kappa \in \mathbb{R}^d$ by solving

$$\mathcal{H}(f \circ \mu_x)(0)\kappa = \nabla(f \circ \mu_x)(0)$$

state of the art



state of the art

Let $x^* \in \mathcal{M}$ be a nondegenerate critical point. Let $\{\mu_x\}_{x\in\mathcal{M}}$ and $\{\nu_x\}_{x\in\mathcal{M}}$ be locally smooth around x^*. Consider the following iteration on $\mathcal{M}$

$$x_0 \in \mathcal{M},\ x_{k+1} = \nu_{x_k}\left(N_{f\circ\mu_{x_k}}(0)\right) \qquad \text{(N)}$$

Theorem: (Hüper-T.) Under the condition

$$\mathrm{D}\,\mu_{x^*}(0) = \mathrm{D}\,\nu_{x^*}(0)$$

there exists an open neighborhood $V \subset \mathcal{M}$ of x^* such that the point sequence generated by (N) converges quadratically to x^* provided $x_0 \in V$.

state of the art

- **know how to construct computable families of coordinate charts for St, Grass**
- **can deal with approximate Newton**
- **local convergence theory for more general iterations (Manton-T.)**
- **some global convergence results of trust region on manifold schemes**

trust-region methods

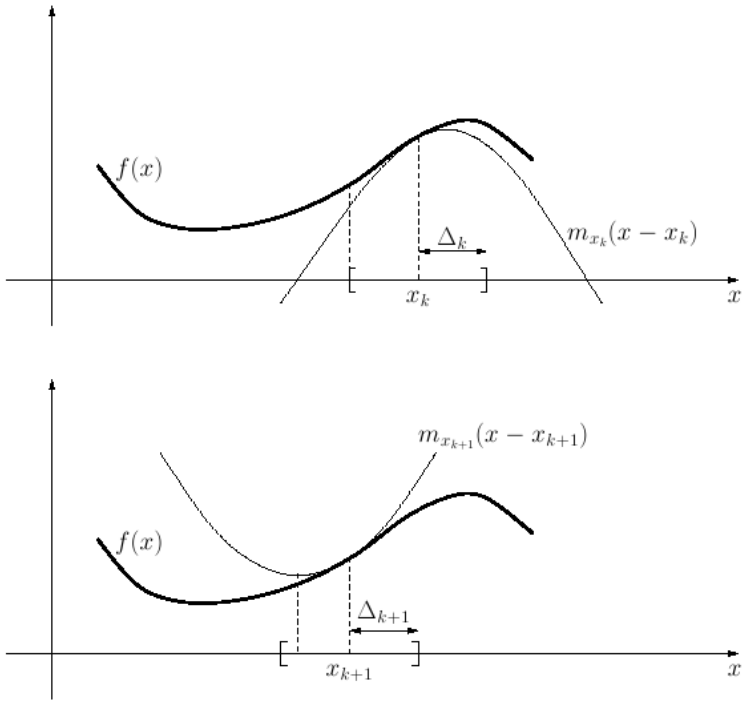


Image: http://www.inma.ucl.ac.be/~blondel/workshops/2004/Absil.pdf

state of the art – ICA

One-unit ICA problem as an optimisation problem on S^{d-1}

$$f : S^{d-1} \to \mathbb{R}, \quad q \mapsto \mathbb{E}[G(q^\top Y)].$$

Geodesics, gradient, Hessian (Hüper-Shen)

$$\gamma_q : \mathbb{R} \to S^{d-1}, \quad \varepsilon \mapsto \exp\left(\varepsilon(\xi q^\top - q\xi^\top)\right) q.$$

$$\operatorname{grad} f(q) = \left(I - qq^\top\right) \mathbb{E}[G'(q^\top Y)\, Y]$$

$$\operatorname{hess} f(q) \cdot \xi = \Big(\underbrace{\mathbb{E}[G''(q^\top Y) Y Y^\top]}_{\in \mathbb{R}^{d\times d}} - \underbrace{\mathbb{E}[G'(q^\top Y) q^\top Y]}_{\in \mathbb{R}}\, I \Big) \cdot \xi$$

state of the art – ICA

THE AUSTRALIAN NATIONAL UNIVERSITY

Alternative to geodesics on S^{d-1}

$$\rho_q : \mathbb{R} \to S^{d-1}, \quad \varepsilon \mapsto \frac{q + \varepsilon\xi}{\|q + \varepsilon\xi\|}$$

ANICA as a selfmap on S^{d-1}

$$q \mapsto \frac{\frac{1}{\tau(q)}(\mathbb{E}[G'(q^\top Y)Y] - \mathbb{E}[G''(q^\top Y)]q)}{\|\frac{1}{\tau(q)}(\mathbb{E}[G'(q^\top Y)Y] - \mathbb{E}[G''(q^\top Y)]q)\|},$$

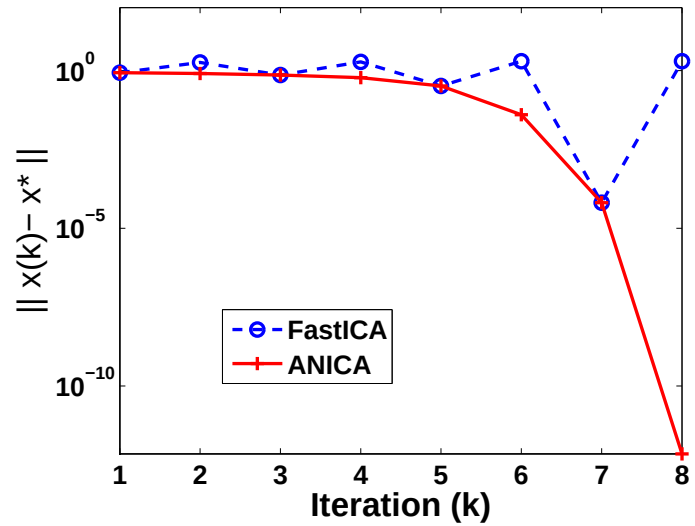
where

$$\tau : S^{d-1} \mapsto \mathbb{R}, \quad \tau(q) := \mathbb{E}[G'(q^\top Y)q^\top Y] - \mathbb{E}[G''(q^\top Y)]$$

NATIONAL ICT AUSTRALIA

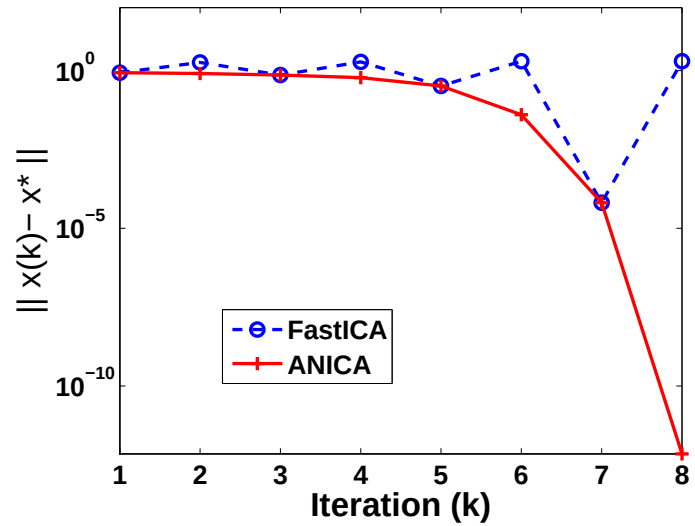
state of the art – ICA

FastICA vs ANICA



state of the art – ICA

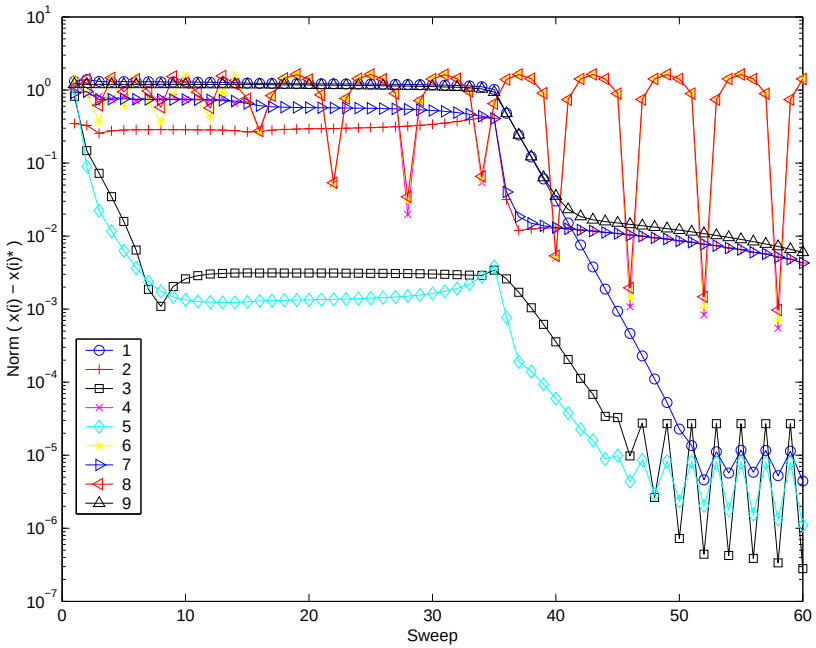
FastICA vs ANICA

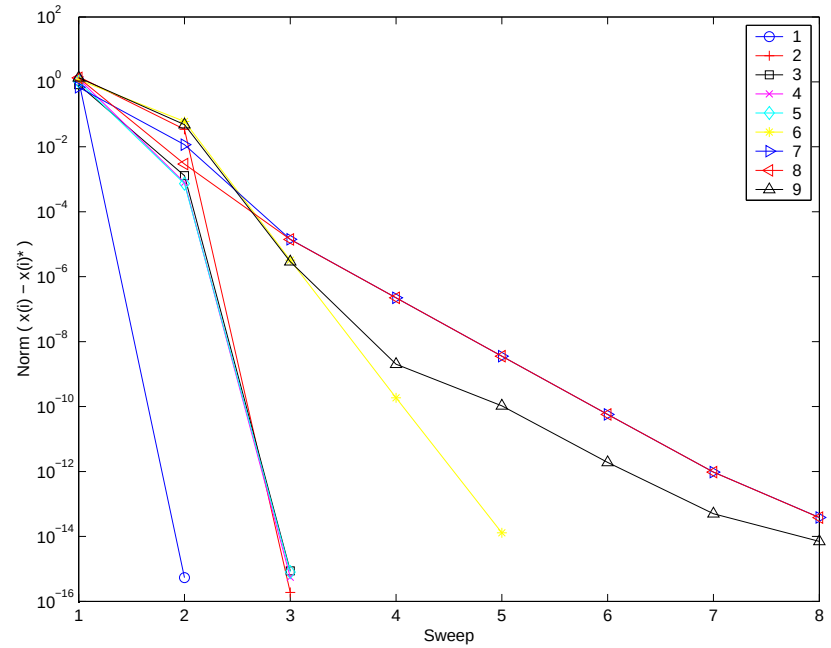


Parallel version (ANLICA, Hüper-Shen) with cost function

$$f : \mathrm{O}_d(\mathbb{R}) \longrightarrow \mathbb{R}, \qquad Q \longmapsto \sum_{i=1}^{m} \mathbb{E}[G(q_i^\top Y)]$$

state of the art – ICA





Parallel FastICA **ANLICA**

the end

THE AUSTRALIAN NATIONAL UNIVERSITY

Thank you.

NATIONAL ICT AUSTRALIA