Nonlinear Signal Processing
(2004-2005)

Course Overview

Instituto Superior Técnico, Lisbon, Portugal

João Xavier

{jxavier}@isr.ist.utl.pt

Outline

- **Motivation**: Signal Processing & Related Applications of Differential Geometry
  - Optimization
  - Kendall's theory of shapes
  - Random Matrix Theory
    - Coherent Capacity of Multi-Antenna Systems
  - Information Geometry
  - Geometrical Interpretation of Jeffreys' Prior
  - Performance Bounds for Constrained or Non-Identifiable Parametric Estimation

- **Course's Table of Contents**
  - Topological manifolds
  - Differentiable manifolds
  - Riemannian manifolds

Outline

- **Bibliography**
  - Recommended textbooks
  - Additional material (short notes on specialized topics)
- **Grading**
- **Discussion, questions, etc**

Applications of DG: Optimization

☐ **Unconstrained minimization problem:**

$$\boldsymbol{x}^* = \arg \min_{\boldsymbol{x} \in \mathbb{R}^n} \; f(\boldsymbol{x})$$

☐ **Iterative line search:**

given initial point $\boldsymbol{x}_0$

for $k = 0, 1, \ldots$

choose descent direction $\boldsymbol{d}_k$

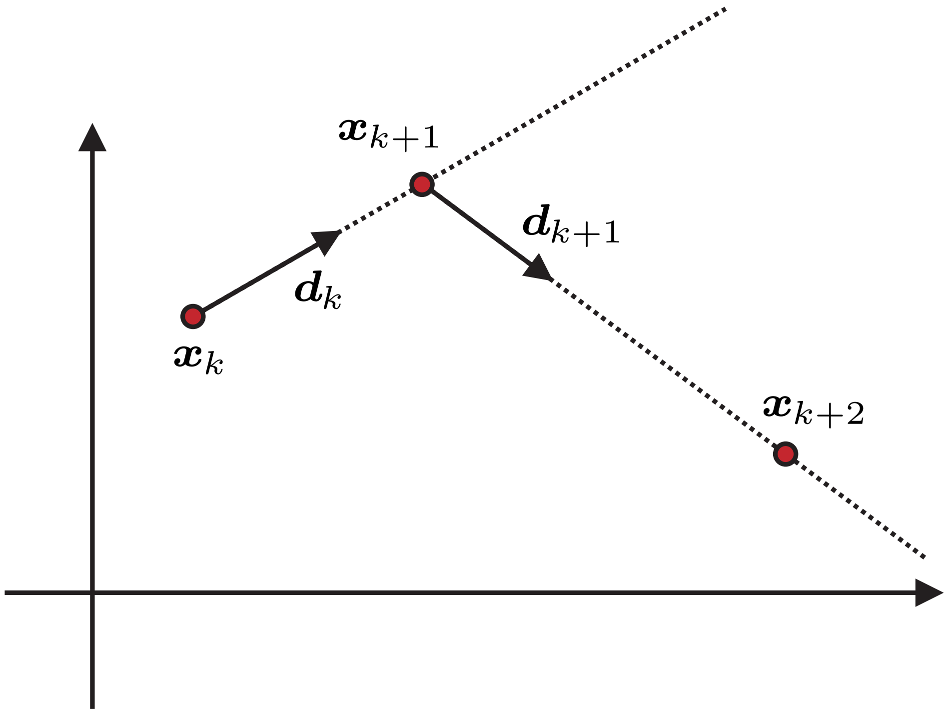
solve $t^* = \arg\min_{t \geq 0} \; f(\boldsymbol{x}_k + t\boldsymbol{d}_k)$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t^* \boldsymbol{d}_k$

end

Applications of DG: Optimization

☐ **Sketch:**



☐ **Descent direction**: $\boldsymbol{d}_{\text{grad}} = -\boldsymbol{\nabla} f(\boldsymbol{x}_k)$, $\boldsymbol{d}_{\text{newton}} = -\left[\boldsymbol{\nabla}^2 f(\boldsymbol{x}_k)\right]^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}_k)$

Applications of DG: Optimization

☐ **Constrained minimization problem:**

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0}} f(\boldsymbol{x})$$

☐ **Iterative line search with projected gradient:**

given initial point $\boldsymbol{x}_0$

for $k = 0, 1, \ldots$

compute $\boldsymbol{d}_k = \mathbf{\Pi}\left(-\boldsymbol{\nabla} f(\boldsymbol{x}_k)\right)$

solve $t^* = \arg\min_{t\geq 0} f(\boldsymbol{x}_k + t\boldsymbol{d}_k)$

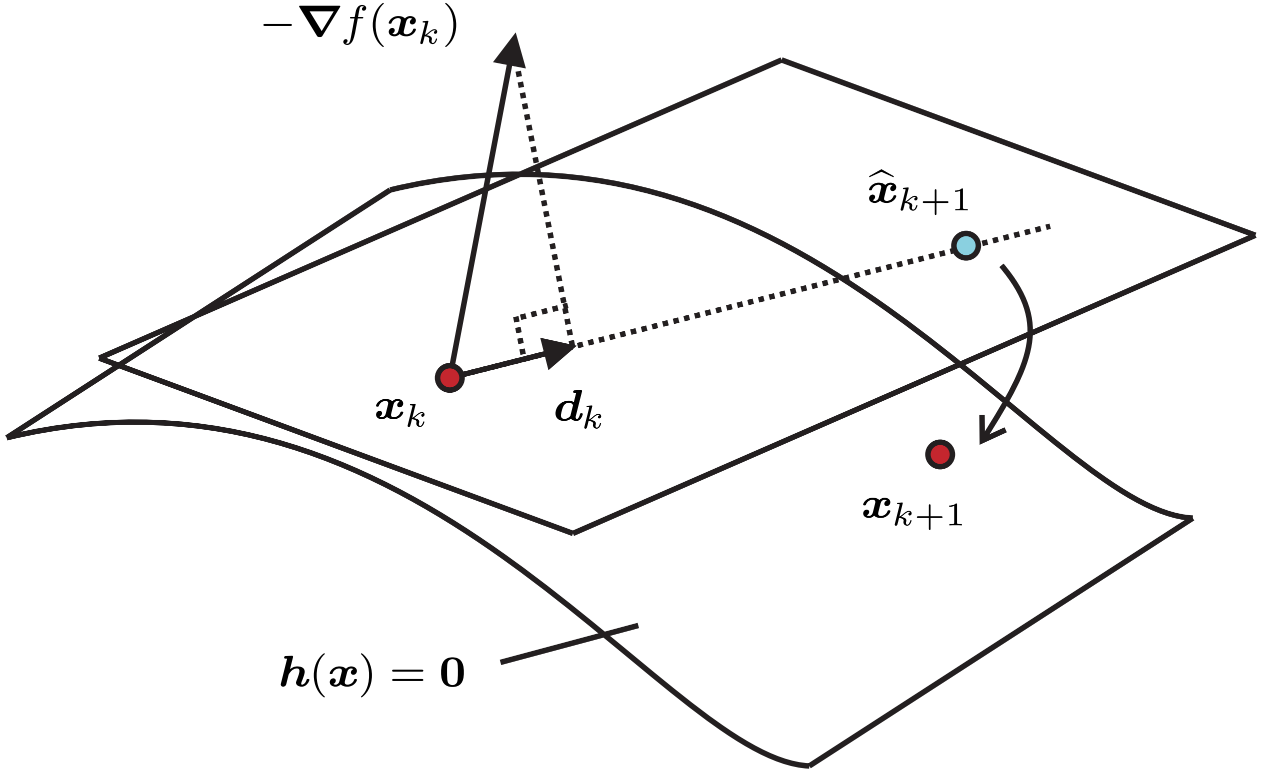
$\widehat{\boldsymbol{x}}_{k+1} = \boldsymbol{x}_k + t^*\boldsymbol{d}_k$

return to the constraint surface $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{h}(\boldsymbol{x})=0} \|\boldsymbol{x} - \widehat{\boldsymbol{x}}_{k+1}\|^2$

end

Applications of DG: Optimization

☐ **Sketch:**



Applications of DG: Optimization

☐ Differential geometry enables a descent algorithm with feasible iterates

☐ **Iterative geodesic search:**

given initial point $\boldsymbol{x}_0$

for $k = 0, 1, \ldots$

choose descent direction $\boldsymbol{d}_k$

solve $t^* = \arg\min_{t \geq 0} \; f(\boldsymbol{\gamma}_k(t))$

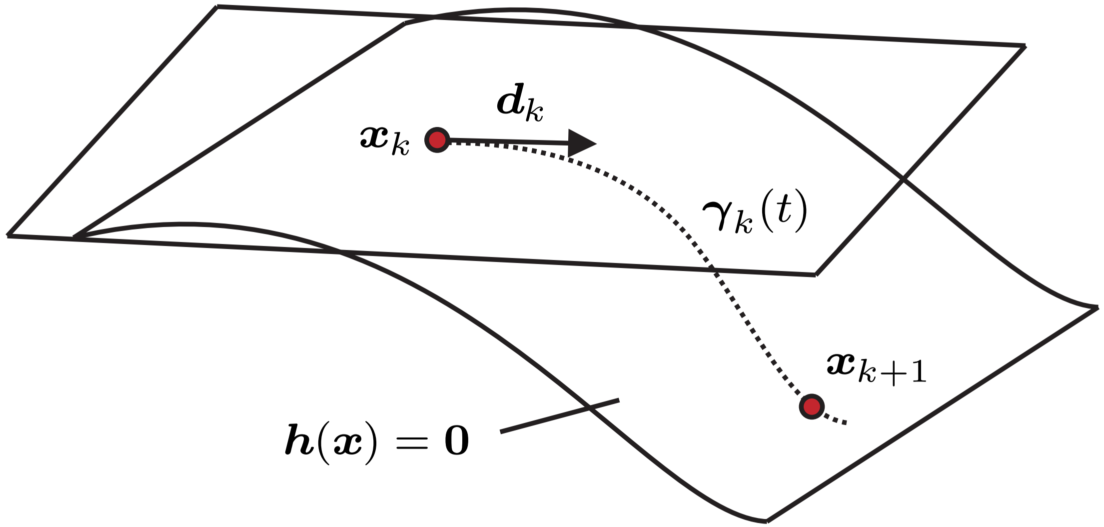
($\boldsymbol{\gamma}_k(t)$= geodesic emanating from $\boldsymbol{x}_k$ in the direction $\boldsymbol{d}_k$)

$\boldsymbol{x}_{k+1} = \boldsymbol{\gamma}_k(t^*)$

end

Applications of DG: Optimization

☐ **Sketch:**



☐ **Descent direction**: generalizations of $\boldsymbol{d}_{\text{grad}}$ and $\boldsymbol{d}_{\text{newton}}$ are available

☐ Theory works for abstract spaces (e.g. projective spaces)

Applications of DG: Optimization

☐ **Example:** Signal model

$$\boldsymbol{y}[t] = \boldsymbol{Q}\boldsymbol{x}[t] + \boldsymbol{w}[t] \qquad t = 1, 2, \ldots, T$$

$\boldsymbol{Q}$: orthogonal matrix ($\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_N$), $\boldsymbol{x}[t]$: known and $\boldsymbol{w}[t] \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{C})$

☐ **Maximum-Likelihood Estimate:**

$$\boldsymbol{Q}^* = \arg \max_{\boldsymbol{Q} \in \mathbb{O}(N)} p\left(\boldsymbol{Y}\,;\, \boldsymbol{Q}\right)$$

- ▷ $\mathbb{O}(N)$= group of $N \times N$ orthogonal matrices
- ▷ $\boldsymbol{Y} = [\, \boldsymbol{y}[1]\, \boldsymbol{y}[2] \, \cdots \, \boldsymbol{y}[T]]$ and $\boldsymbol{X} = [\, \boldsymbol{x}[1]\, \boldsymbol{x}[2] \, \cdots \, \boldsymbol{x}[T]]$

Applications of DG: Optimization

☐ **Optimization problem:** Orthogonal Procrustes rotation

$$\begin{aligned} \boldsymbol{Q}^* &= \arg \min_{\boldsymbol{Q} \in \mathbb{O}(N)} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|^2_{\boldsymbol{C}^{-1}} \\ &= \arg \min_{\boldsymbol{Q} \in \mathbb{O}(N)} \mathrm{tr}\left\{\boldsymbol{Q}^T \boldsymbol{C}^{-1} \boldsymbol{Q} \widehat{\boldsymbol{R}}_{\boldsymbol{xx}}\right\} - \mathrm{tr}\left\{\boldsymbol{Q}^T \boldsymbol{C}^{-1} \widehat{\boldsymbol{R}}_{\boldsymbol{yx}}\right\} \end{aligned}$$

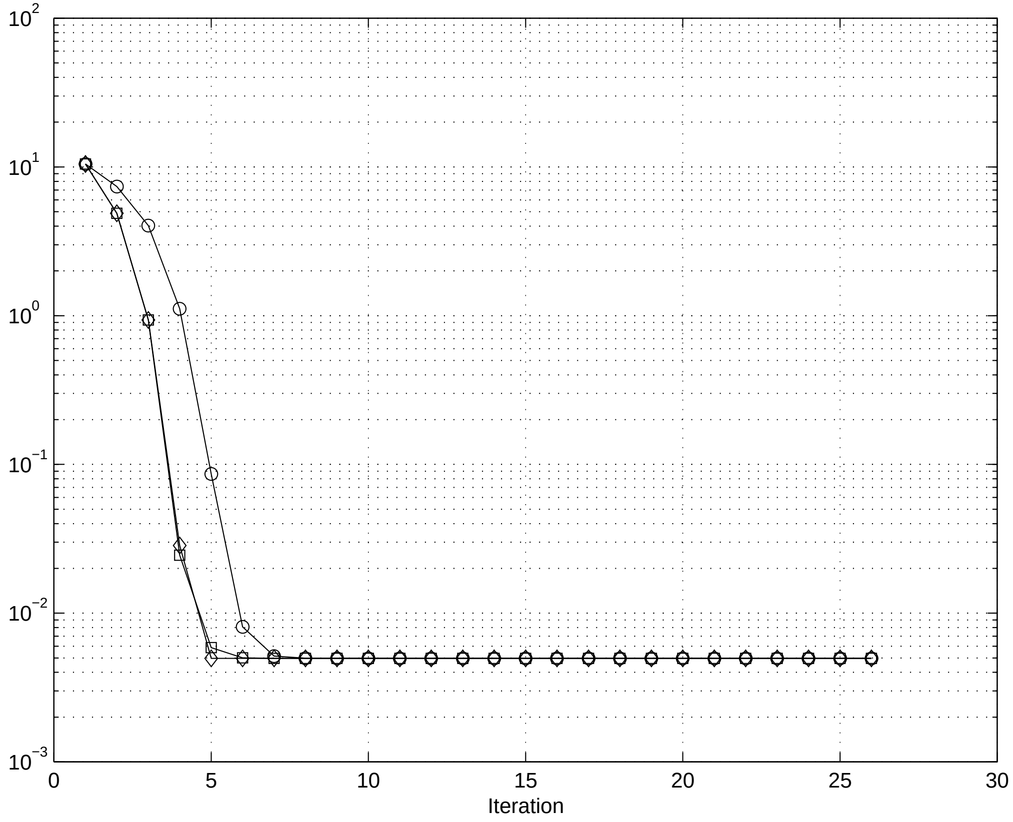
▷ $\widehat{\boldsymbol{R}}_{\boldsymbol{yx}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{y}[t]\boldsymbol{x}[t]^T$ and $\widehat{\boldsymbol{R}}_{\boldsymbol{xx}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}[t]\boldsymbol{x}[t]^T$

☐ **Note:** the eigenstructure of $\boldsymbol{C}$ controls the Hessian of the objective

$\kappa(\boldsymbol{C}^{-1}) = \dfrac{\lambda_{\mathsf{max}}(\boldsymbol{C}^{-1})}{\lambda_{\mathsf{min}}(\boldsymbol{C}^{-1})}$ condition number of $\boldsymbol{C}^{-1}$

Applications of DG: Optimization

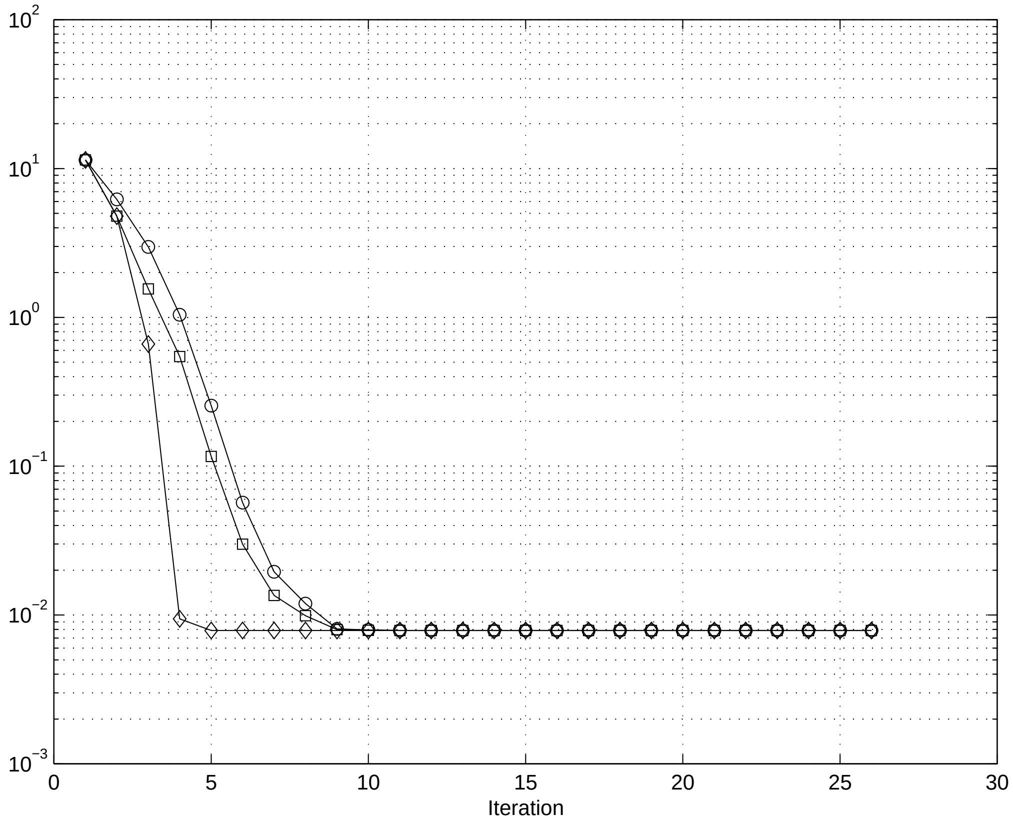
☐ **Example:** $N = 5$, $T = 100$, $\boldsymbol{C} = \mathrm{diag}(1, 1, 1, 1, 1)$, $\kappa(\boldsymbol{C}^{-1}) = 1$



○=projected gradient □=gradient geodesic descent ◇=Newton geodesic descent

Applications of DG: Optimization

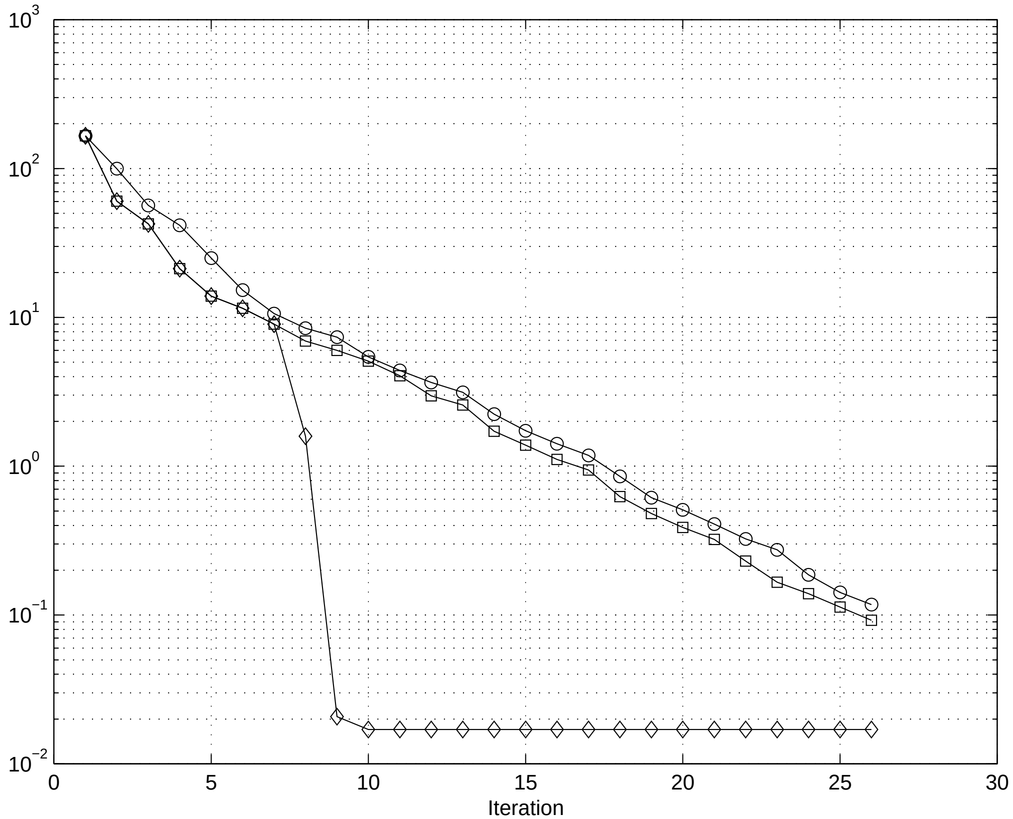
☐ **Example:** $N = 5$, $T = 100$, $\boldsymbol{C} = \text{diag}(0.2, 0.4, 0.6, 0.8, 1)$, $\kappa(\boldsymbol{C}^{-1}) = 5$



○=projected gradient ☐=gradient geodesic descent ◇=Newton geodesic descent

Applications of DG: Optimization

☐ **Example:** $N = 5$, $T = 100$, $\boldsymbol{C} = \text{diag}(0.02, 0.05, 0.14, 0.37, 1)$, $\kappa(\boldsymbol{C}^{-1}) = 50$



$\circ$=projected gradient $\square$=gradient geodesic descent $\diamond$=Newton geodesic descent

Applications of DG: Optimization

☐ **Important:** Following geodesics is not necessarily optimal. See:

"Optimization algorithms exploiting unitary constraints", J. Manton, IEEE Trans. on Signal Processing, vol. 50, no. 3, pp. 635–650, March 2002

Applications of DG: Optimization

- **Bibliography:**

- "The geometry of weighted low-rank approximations", J. Manton *et al.*, IEEE Trans. on Signal Processing, vol. 51, no. 2, pp. 500–514, February 2003
- "Efficient algorithms for inferences on Grassmann manifolds", K. Gallivan *et al*, Proc. 12th IEEE Workshop Statistical Signal Processing, 2003
- "Adaptive eigenvalue computations using Newton's method on the Grassmann manifold", E. Lundstrom *et al.*, SIAM J. Matrix Anal. Appl., vol. 23, no. 3, pp. 819–839, 2002
- "A Grassmann-Rayleigh quotient iteration for computing invariant subspaces", P. Absil *et al.*, SIAM Review, vol. 44, no. 1, pp. 57–73, 2002
- "Algorithms on the Stiefel manifold for joint diagonalization", M. Nikpour *et al.*, IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP), vol. 2, pp. 1481–1484, 2002
- "Optimization algorithms exploiting unitary constraints", J. Manton, IEEE Trans. on Signal Processing, vol. 50, no. 3, pp. 635–650, March 2002
- "Contravariant adaptation on structured matrix spaces", T. Moon and J. Gunther, Signal Processing, 82, pp. 1389–1410, 2002

Applications of DG: Optimization

☐ **Bibliography (cont.):**

- "The geometry of the Newton method on non-compact Lie groups", R. Mahony and J. Manton, Journal of Global Optimization, vol. 23, pp. 309–327, 2002.
- "Prior knowledge and preferential structures in gradient descent learning algorithms", R. Mahony and Williamson, Journal of Machine Learning Research, pp. 311–355, 2001.
- "Precoder assisted channel estimation in complex projective space", J. Manton, IEEE 3rd Workshop on Sig. Proc. Advanc. on Wir. Comm. (SPAWC), pp. 348–351, 2001
- "Optimization on Riemannian manifold", IEEE Proc. 38th conference on Decision and Control, pp. 888–893, Dec. 1999.
- "Optimum phase-only adaptive nulling", S. Smith, IEEE Trans. on Signal Processing, vol. 47, no. 7, pp. 1835–1843, July 1999
- "Motion estimation in computer vision: optimization on Stiefel manifolds", Y. Ma *et al*, IEEE Proc. 38th conference on Decision and Control, vol. 4, pp. 3751–3756, Dec. 1998
- "The geometry of algorithms with orthogonality constraints", A. Edelman *et al.*, SIAM J. Matrix Anal. Appl., vol. 20, no. 2, pp. 303–353, 1998

Applications of DG: Optimization

□ **Bibliography (cont.):**

◇ "Optimal motion from image sequences: a Riemannian viewpoint", Y. Ma *et al*, Electronic Research Lab Memorandum, UC Berkeley, 1998

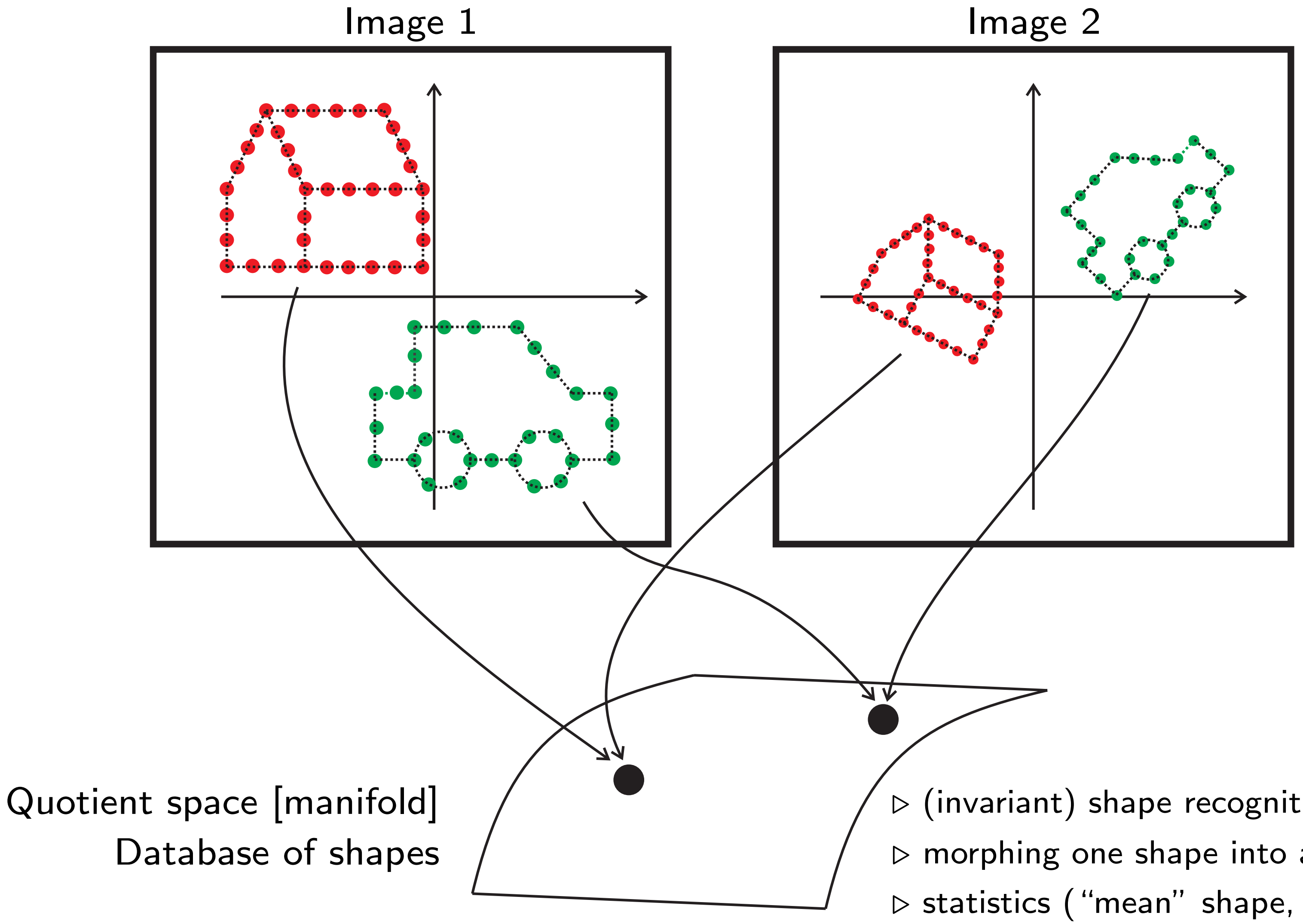
◇ "Optimization tecnhiques on Riemannian manifolds", S. Smith, Fields Institute Communications, vol. 3, pp. 113–136, 1994

◇ "Optimization and Dynamical Systems", U. Helmke and J. Moore, Springer-Verlag, 1994

◇ "Geometric optimization methods for adaptive filtering", S. Smith, PhD Thesis, Harvard University, 1993

◇ "Constrained optimization along geodesics", C. Botsaris, J. Math. Anal. Appl., vol. 79, pp. 295–306, 1981

Applications of DG: Kendall's theory of shapes



- (invariant) shape recognition
- morphing one shape into another
- statistics ("mean" shape, clustering)

Applications of DG: Kendall's theory of shapes

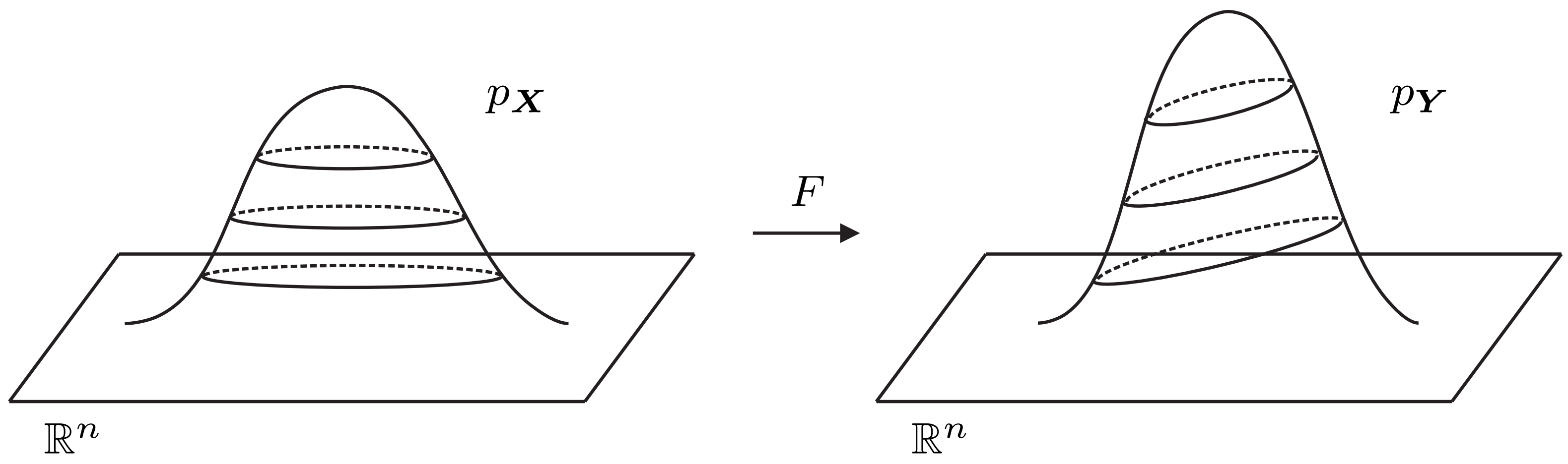
- **Bibliography:**

  - "Multivariate shape analysis", I. Dryden and K. Mardia, Sankhya: The Indian Journal of Statistics, 55, pp. 460–480, 1993
  - "Procrustes methods in the statistical analysis of shape", C. Goodall, J. R. Statist. Soc. B, 53, no.2, pp. 285–339, 1991
  - "A survey of the statistical theory of shapes", D. Kendall, Statist. Sci., 4, pp. pp. 87–120, 1989
  - "Shape manifolds, Procrustean metrics and complex projective spaces", D. Kendall, Bull. London Math. Soc., 16, pp. 81–121, 1984
  - "Directional Statistics", K. Mardia and P. Jupp, Wiley Series in Probability and Statistics

Applications of DG: Random Matrix Theory

☐ **Basic statistics:** transformation of random objects in Euclidean spaces

$$\begin{cases} \boldsymbol{x} \text{ is a random vector in } \mathbb{R}^n \\ \boldsymbol{x} \sim p_{\boldsymbol{X}}(\boldsymbol{x}) \\ F : \mathbb{R}^n \to \mathbb{R}^n \text{ smooth, bijective} \\ \boldsymbol{y} = F(\boldsymbol{x}) \end{cases} \quad \Rightarrow \quad \begin{array}{l} \boldsymbol{y} \sim p_{\boldsymbol{Y}}(\boldsymbol{y}) = p_{\boldsymbol{X}}(F^{-1}(\boldsymbol{y}))\, J(\boldsymbol{y}) \\ J(\boldsymbol{y}) = \dfrac{1}{\det(DF(F^{-1}(\boldsymbol{y})))} \end{array}$$

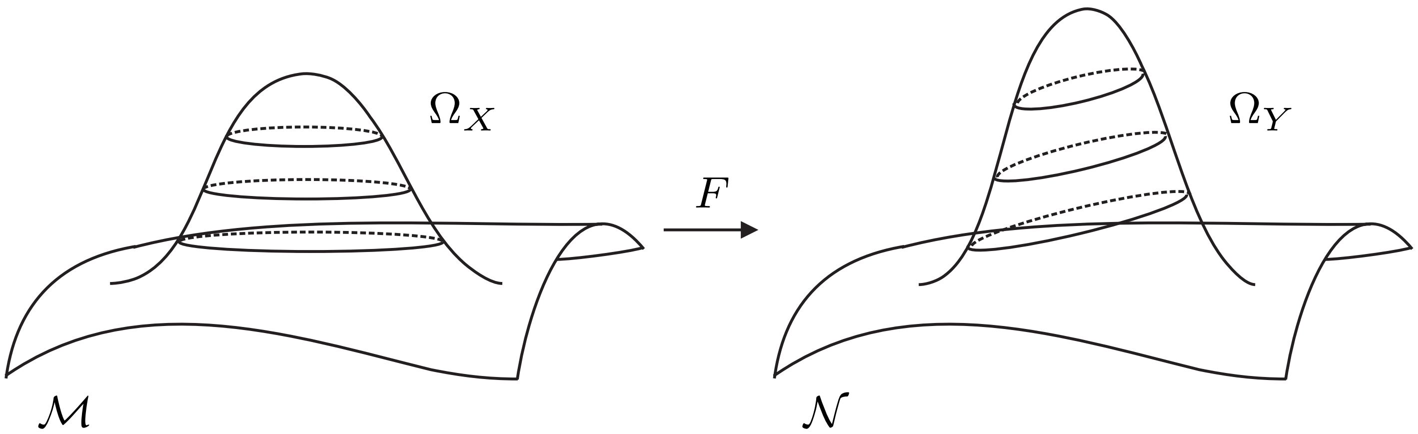


Applications of DG: Random Matrix Theory

☐ **Generalization:** transformation of random objects in manifolds $\mathcal{M}, \mathcal{N}$

$$\left\{\begin{array}{l} x \text{ is a random point in } \mathcal{M} \\ x \sim \Omega_X \text{ (exterior form)} \\ F : \mathcal{M} \to \mathcal{N} \text{ smooth, bijective} \\ y = F(x) \end{array}\right. \quad \Rightarrow \quad y \sim \Omega_Y = \ldots$$

The answer is provided by the calculus of exterior differential forms



Applications of DG: Random Matrix Theory

☐ **Example A:** decoupling a random vector in amplitude and direction

$$\left\{\begin{array}{l} \mathcal{M} = \mathbb{R}^n - \{\mathbf{0}\} = \{\boldsymbol{x} \,:\, \boldsymbol{x} \neq \mathbf{0}\} \\ \mathcal{N} = \mathbb{R}^+ \times \mathbb{S}^{n-1} = \{(R, \boldsymbol{u}) \,:\, R > 0, \|\boldsymbol{u}\| = 1\} \\ (R, \boldsymbol{u}) = F(\boldsymbol{x}) = \left(\|\boldsymbol{x}\|\,, \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right) \\ \boldsymbol{x} \sim p_{\boldsymbol{X}}(\boldsymbol{x}) \end{array}\right. \quad \Rightarrow \quad p(R, \boldsymbol{u}) = p_{\boldsymbol{X}}(R\boldsymbol{u})\, R^{n-1}$$

☐ **Example B:** decoupling a random matrix through the polar decomposition

$$\left\{\begin{array}{l} \mathcal{M} = \mathbb{GL}(n) = \left\{\boldsymbol{X} \in \mathbb{R}^{n \times n} \,:\, |\boldsymbol{X}| \neq \mathbf{0}\right\} \\ \mathcal{N} = \mathbb{P}(n) \times \mathbb{O}(n) = \left\{(\boldsymbol{P}, \boldsymbol{Q}) \,:\, \boldsymbol{P} \succ \mathbf{0}, \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n\right\} \\ (\boldsymbol{P}, \boldsymbol{Q}) = F(\boldsymbol{X}) \Leftrightarrow \boldsymbol{X} = \boldsymbol{P}\boldsymbol{Q} \\ \boldsymbol{X} \sim p_{\boldsymbol{X}}(\boldsymbol{X}) \end{array}\right. \quad \Rightarrow \quad p(\boldsymbol{P}, \boldsymbol{Q}) = \ldots \text{(known)}$$

Applications of DG: Random Matrix Theory

☐ **Example C:** decoupling a random symmetric matrix by eigendecomposition

$$\left\{\begin{array}{l} \mathcal{M} = \mathbb{S}(n) = \left\{ \boldsymbol{X} \in \mathbb{R}^{n\times n} \,:\, \boldsymbol{X} = \boldsymbol{X}^T \right\} \\ \mathcal{N} = \mathbb{O}(n) \times \mathbb{D}(n) = \left\{ (\boldsymbol{Q}, \boldsymbol{\Lambda}) \,:\, \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n, \boldsymbol{\Lambda} : \text{diag} \right\} \\ (\boldsymbol{Q}, \boldsymbol{\Lambda}) = F(\boldsymbol{X}) \Leftrightarrow \boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \\ \boldsymbol{X} \sim p_{\boldsymbol{X}}(\boldsymbol{X}) \end{array}\right. \quad \Rightarrow \quad p(\boldsymbol{Q}, \boldsymbol{\Lambda}) = \ldots \text{(known)}$$

☐ **Many other examples...** (e.g. Cholesky, QR, LU, SVD)

Applications of DG: Random Matrix Theory

Bibliography:

- "Matrix Variate Distributions", A. Gupta, Chapman & Hall, 1999
- "Jacobians of Matrix Transformations and Functions of Matrix Argument", A. Mathai, World Scientific, 1997
- "Random Matrices", M. Mehta, Academic Press, 1991
- "Eigenvalues and Condition Numbers of Random Matrices", A. Edelman, PhD Thesis, Massachusetts Institute of Technology, 1989
- "Multivariate Calculation", R. Farrell, Springer-Verlag, 1985
- "Aspects of Multivariate Statistical Theory", R. Muirhead, John Wiley & Sons, 1982
- "Distributions of matrix variates and latent roots derived from normal samples", A. James, Annals of Math. Statistics, vol. 35, pp. 475–501, 1964

RMT and DG concepts in signal processing

□ **Bibliography (only a small sample):**

◇ "Random Matrix Theory and Wireless Communications", A. Tulino and S. Verdú, Now Publishers Inc., 2004.

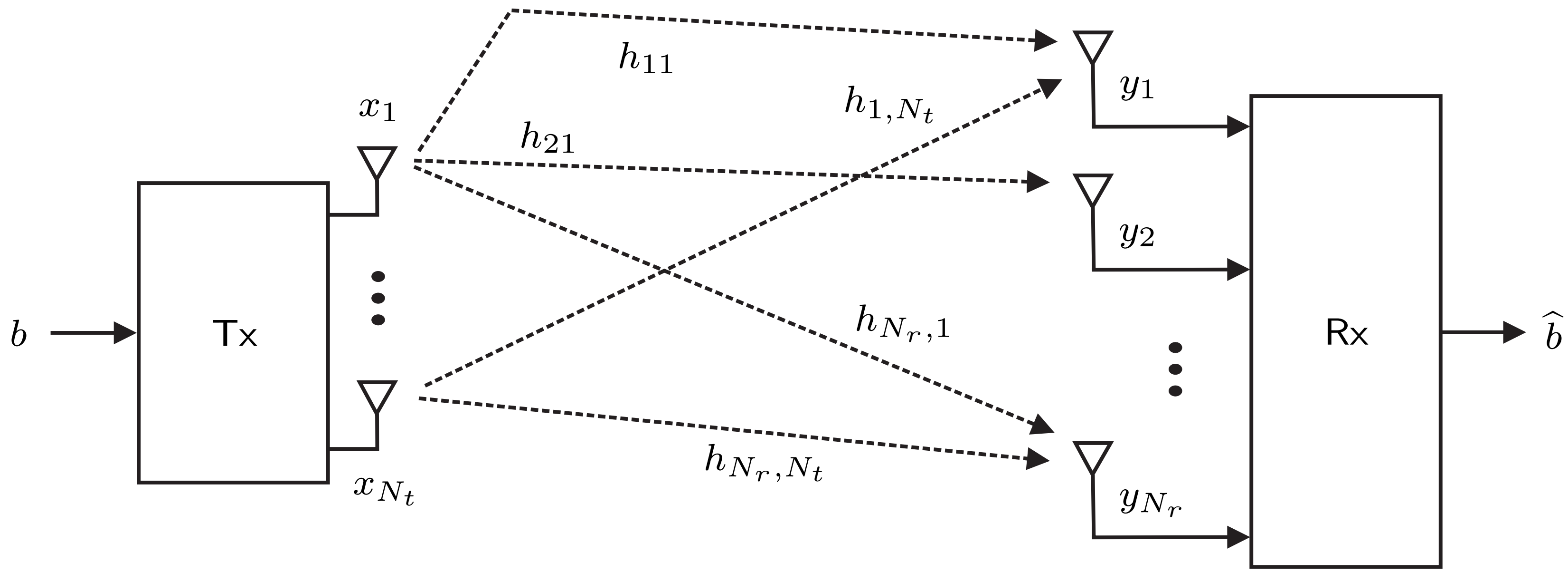
◇ "Grassmann-based signal design for non-coherent reception", I. Kammoun and J. C. Belfiore, *Signal Processing Advances in Wireless Communications, 2003*, SPAWC 2003, 4th IEEE Workshop 2003 (pp.507–511)

◇ "Communication on the Grassmann manifold: a geometric approach to the nonchoerent multiple-antenna channel", L. Zheng and D. Tse, *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, February 2002.

◇ A. Srivastava, "A Bayesian approach to geometric subspace estimation," *IEEE Transactions on Signal Processing*, vol. 48, no. 5, pp. 1390–1400, May 2000.

Applications of RMT: Coherent Capacity of Multi-Antenna Systems

☐ **Scenario:** point-to-point single-user communication with multiple Tx antennas



Applications of RMT: Coherent Capacity of Multi-Antenna Systems

☐ **Data model:** $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ with $\boldsymbol{y}, \boldsymbol{n} \in \mathbb{C}^{N_r}$, $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$, $\boldsymbol{x} \in \mathbb{C}^{N_t}$

- ◇ N_t = number of Tx antennas
- ◇ N_r = number of Rx antennas

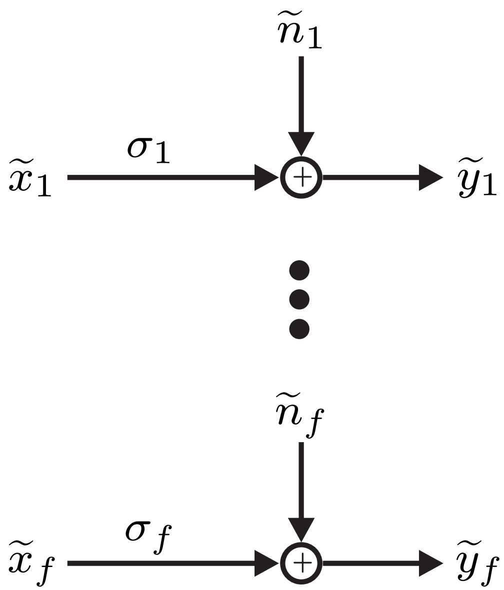
Assumption: $n_i \overset{\text{iid}}{\sim} \mathbb{C}\mathcal{N}(0, 1)$

☐ **Decoupled data model:**

- ◇ SVD: $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ with $\boldsymbol{U} \in \mathbb{U}(N_r)$, $\boldsymbol{V} \in \mathbb{U}(N_t)$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_f, \boldsymbol{0})$, $(\sigma_1, \ldots, \sigma_f)$ = nonzero singular values of $\boldsymbol{H}$, $f = \min\{N_r, N_t\}$
- ◇ Transform the data: $\widetilde{\boldsymbol{y}} = \boldsymbol{U}^H\boldsymbol{y}$, $\widetilde{\boldsymbol{x}} = \boldsymbol{V}^H\boldsymbol{x}$ and $\widetilde{\boldsymbol{n}} = \boldsymbol{U}^H\boldsymbol{n}$
- ◇ Equivalent diagonal model: $\widetilde{\boldsymbol{y}} = \boldsymbol{\Sigma}\widetilde{\boldsymbol{x}} + \widetilde{\boldsymbol{n}}$

Applications of RMT: Coherent Capacity of Multi-Antenna Systems

☐ **Interpretation:** The matrix channel $\boldsymbol{H}$ is equivalent to f parallel scalar channels



Applications of RMT: Coherent Capacity of Multi-Antenna Systems

☐ **Assumption:** $\boldsymbol{H}$ is random and known only at the Rx

☐ **Channel capacity:**

$$C = \max_{p(\boldsymbol{x}), \mathsf{E}\{\|\boldsymbol{x}\|^2 \leq P\}} I(\boldsymbol{x}; (\boldsymbol{y}, \boldsymbol{H}))$$

I = mutual information

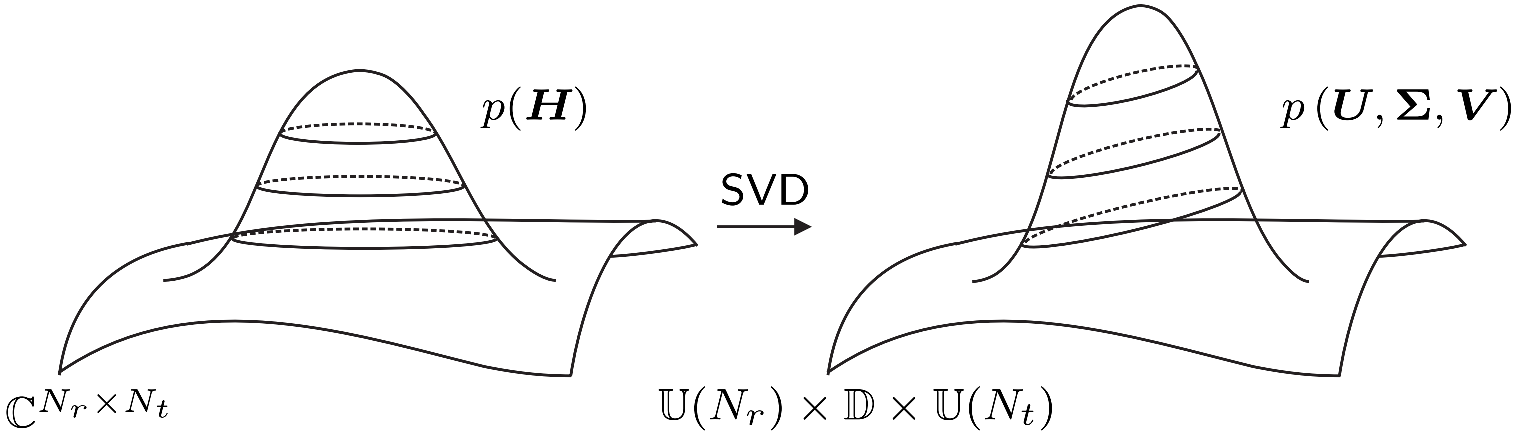
☐ **Solution:**

$$C = \mathsf{E}_{\boldsymbol{H}} \left\{ \sum_{i=1}^{f} \log\left(1 + (P/N_t)\sigma_i^2\right) \right\}$$

Recall: $(\sigma_1, \ldots, \sigma_f)$ = nonzero singular values of $\boldsymbol{H}$, $f = \min\{N_r, N_t\}$

Applications of RMT: Coherent Capacity of Multi-Antenna Systems

☐ $\boldsymbol{H}$ is random and $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ (SVD)



☐ **Capacity:** when $[\boldsymbol{H}_{ij}] \overset{\text{iid}}{\sim} \mathbb{C}\mathcal{N}(0,1)$

$$C = \int_0^\infty \log(1 + (P/N_t)\lambda) \sum_{k=0}^{f-1} \frac{k!}{(k+g-f)!} (L_k^{g-f}(\lambda))^2 \lambda^{g-f} e^{-\lambda}\, d\lambda$$

$g = \max\{N_r, N_t\}$ and L_j^i=Laguerre polynomials

Applications of RMT: Coherent Capacity of Multi-Antenna Systems

□ **Bibliography:**

◇ "Keyholes, correlations and capacities of multielement transmit and receive antennas", D. Chizhik, IEEE Trans. Wireless Comm., vol. 1, pp. 361–368, April 2002

◇ "Capacity scaling in MIMO wireless systems under correlated fading", C. Chuah, IEEE Trans. Information Th., vol. 48, pp. 637–650, March 2002

◇ "Capacity of mobile multiple-antenna communication link in Rayleigh flat-fading", T. Marzetta *et al.*, IEEE Trans. Information Th., vol. 45, no. 1, pp. 139–157, January 1999

◇ "On limits of wireless communications in fading environment when using multiple antennas", G. Foschini and M. Gans, Wireless Personal Communications, vol. 6, no.3, pp. 311–355, 1998

◇ "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas", G. Foschini, Bell Labs Technical Journal, vol. 1, no. 2, pp. 41–59, 1996

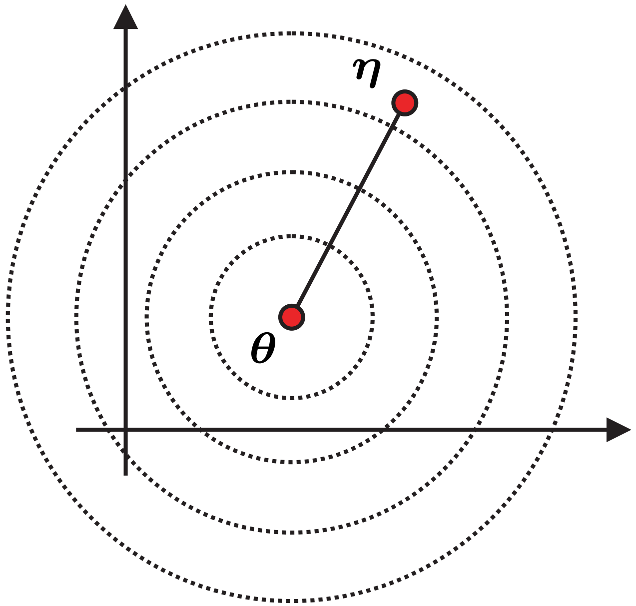
◇ "Capacity of multi-antenna Gaussian channels", I. Telatar, AT&T Bell Labs, Internal Technical Memorandum, 1995

Applications of DG: Information Geometry

☐ **Problem:** Given a parametric statistical family $\mathcal{F} = \{p(\boldsymbol{x};\boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \Theta\}$ assign a distance function $d \,:\, \Theta \times \Theta \to \mathbb{R}$

☐ **Example:** $\mathcal{F} = \{p(\boldsymbol{x};\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \,:\, \boldsymbol{\theta} \in \Theta = \mathbb{R}^n\}$ (note: $\boldsymbol{\Sigma}$ is fixed)

Naive choice (Euclidean distance): $d(\boldsymbol{\theta}, \boldsymbol{\eta}) = \|\boldsymbol{\theta} - \boldsymbol{\eta}\|$



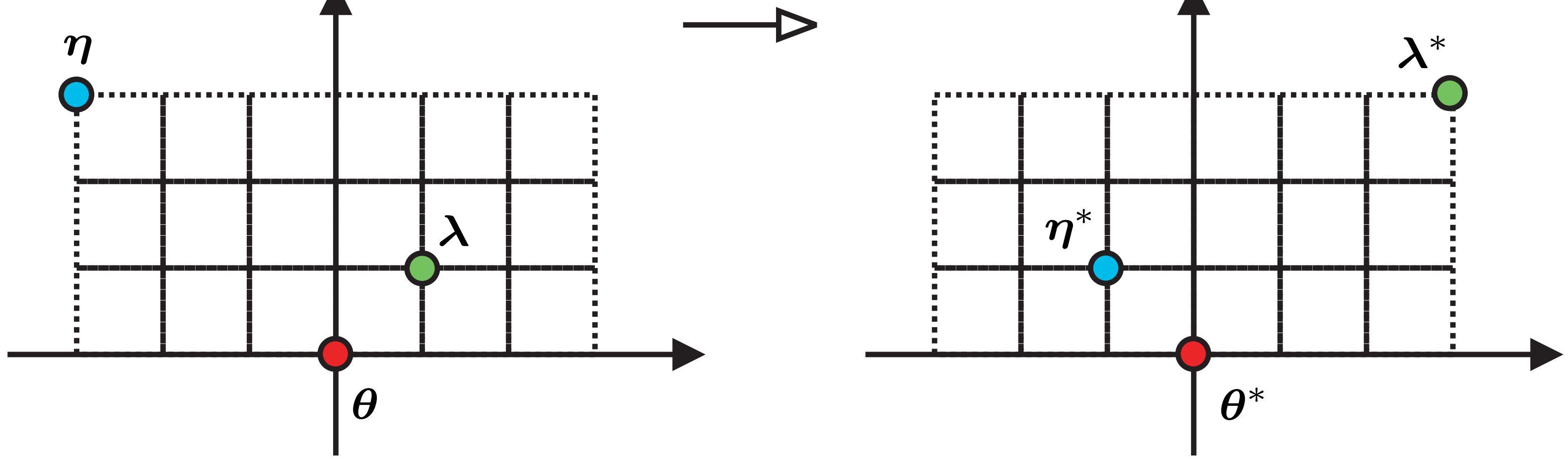
☐ This method does not produce "intrinsic" distances (parameter invariant)

Applications of DG: Information Geometry

☐ With $\boldsymbol{\theta}^* = \boldsymbol{A}\boldsymbol{\theta}$: $\mathcal{F} = \left\{ p(\boldsymbol{x}; \boldsymbol{\theta}^*) \sim \mathcal{N}(\boldsymbol{A}^{-1}\boldsymbol{\theta}^*, \boldsymbol{\Sigma}) \,:\, \boldsymbol{\theta}^* \in \Theta^* = \mathbb{R}^n \right\}$

☐ **Example:** $\boldsymbol{\theta} = (0, 0), \boldsymbol{\eta} = (-3, 3), \boldsymbol{\lambda} = (1, 1), \boldsymbol{A} = \begin{bmatrix} 5/3 & 4/3 \\ 4/3 & 5/3 \end{bmatrix}$

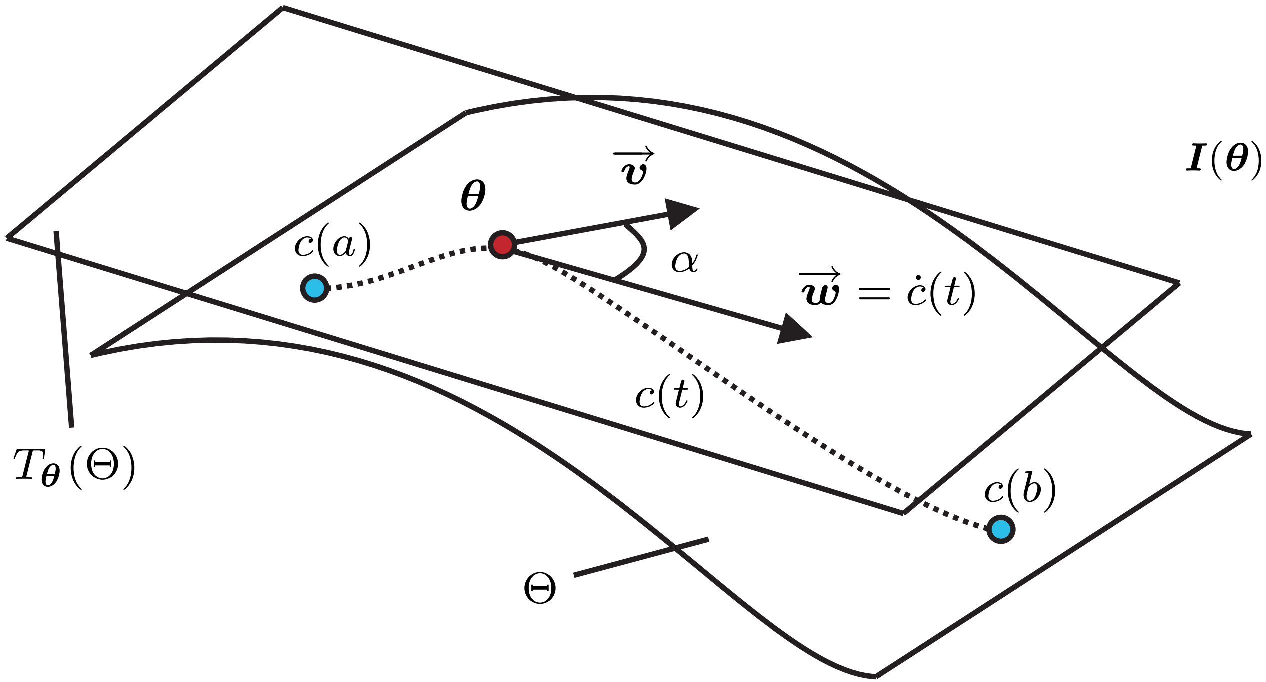
$$\boldsymbol{\theta}^* = \boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{\eta}^* = \boldsymbol{A}\boldsymbol{\eta}, \boldsymbol{\lambda}^* = \boldsymbol{A}\boldsymbol{\lambda}$$



$d(\boldsymbol{\theta}, \boldsymbol{\lambda}) < d(\boldsymbol{\theta}, \boldsymbol{\eta})$ $\qquad$ $d(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*) > d(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$

Applications of DG: Information Geometry

- Rao suggested the information metric to obtain distances between pdf's
- **Differential geometric interpretation:** The Fisher Information Matrix is adopted as the Riemannian tensor on Θ



$$\langle \vec{\boldsymbol{v}}, \vec{\boldsymbol{w}} \rangle = \vec{\boldsymbol{v}}^T \boldsymbol{I}(\boldsymbol{\theta}) \vec{\boldsymbol{w}}$$

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}} \left\{ \boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log p(\boldsymbol{x}; \boldsymbol{\theta}) \right\}$$

$$|\vec{\boldsymbol{v}}| = \sqrt{\langle \vec{\boldsymbol{v}}, \vec{\boldsymbol{v}} \rangle}$$

$$\alpha = \frac{\langle \vec{\boldsymbol{v}}, \vec{\boldsymbol{w}} \rangle}{|\vec{\boldsymbol{v}}| \, |\vec{\boldsymbol{w}}|}$$

$$\text{length(c)} = \int_a^b |\dot{c}(t)| \, dt$$

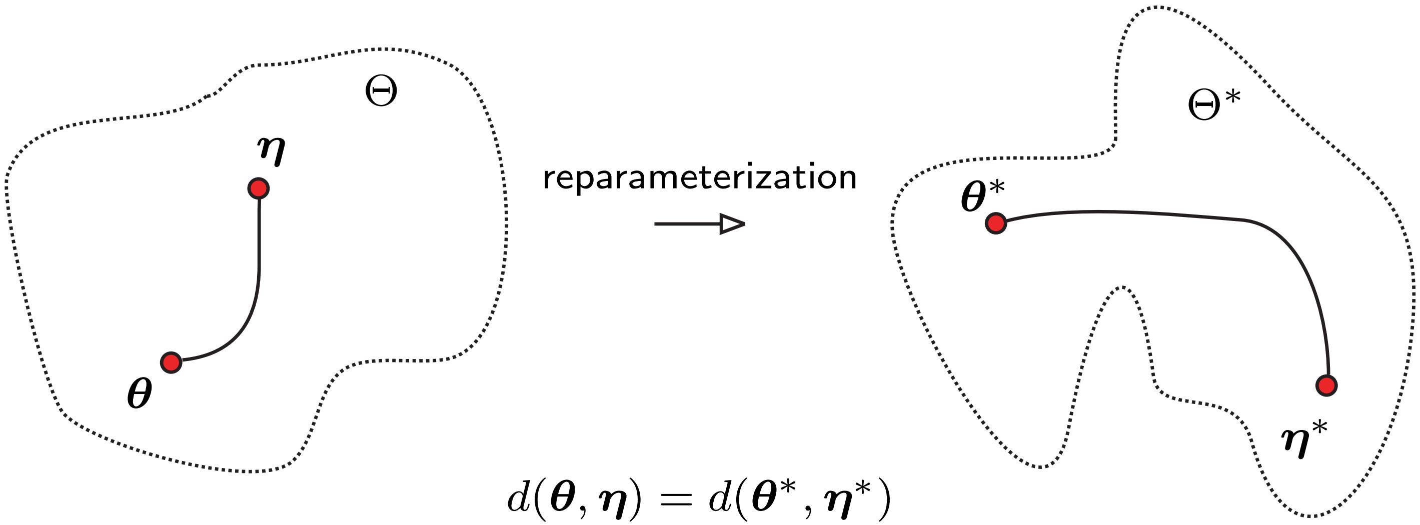
- **Insight:** A parametric statistical family is an autonomous geometrical object

Applications of DG: Information Geometry

☐ **Information distance:**

$$d(\boldsymbol{\theta}, \boldsymbol{\eta}) = \inf \{\text{length}(c) \; : \; c \text{ is a curve on } \Theta \text{ connecting } \boldsymbol{\theta} \text{ to } \boldsymbol{\eta}\}$$

☐ The information distance is invariant to reparameterizations



$$d(\boldsymbol{\theta}, \boldsymbol{\eta}) = d(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$$

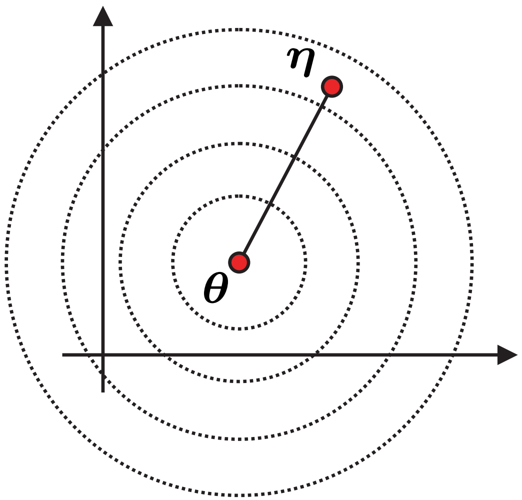
☐ **Link with Kullback-Leibler distance:** $d_{\mathsf{KL}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{2}\, d(\boldsymbol{\theta}, \boldsymbol{\eta})^2 + O\left(d(\boldsymbol{\theta}, \boldsymbol{\eta})^3\right)$

Applications of DG: Information Geometry

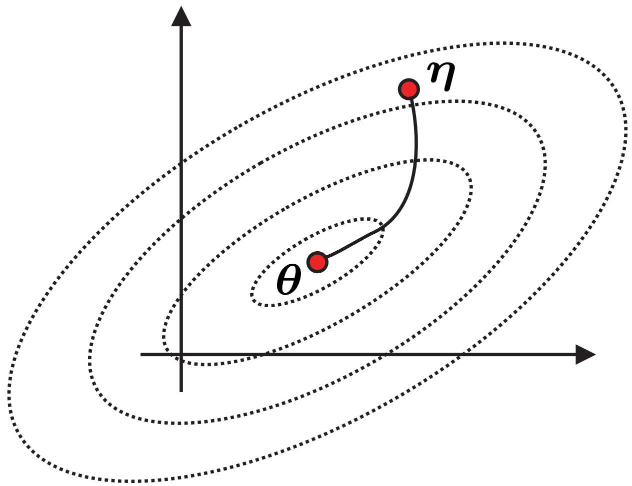
☐ **Some examples:**

$\diamond$ $\mathcal{F} = \{p(\boldsymbol{x};\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta},\boldsymbol{\Sigma}) \,:\, \boldsymbol{\theta} \in \Theta = \mathbb{R}^n\}$ ($\boldsymbol{\Sigma}$ is fixed)

$$d(\boldsymbol{\theta},\boldsymbol{\eta}) = \sqrt{(\boldsymbol{\theta}-\boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}-\boldsymbol{\eta})}$$ [Mahalanobis distance]



Euclidean distance (geodesic)

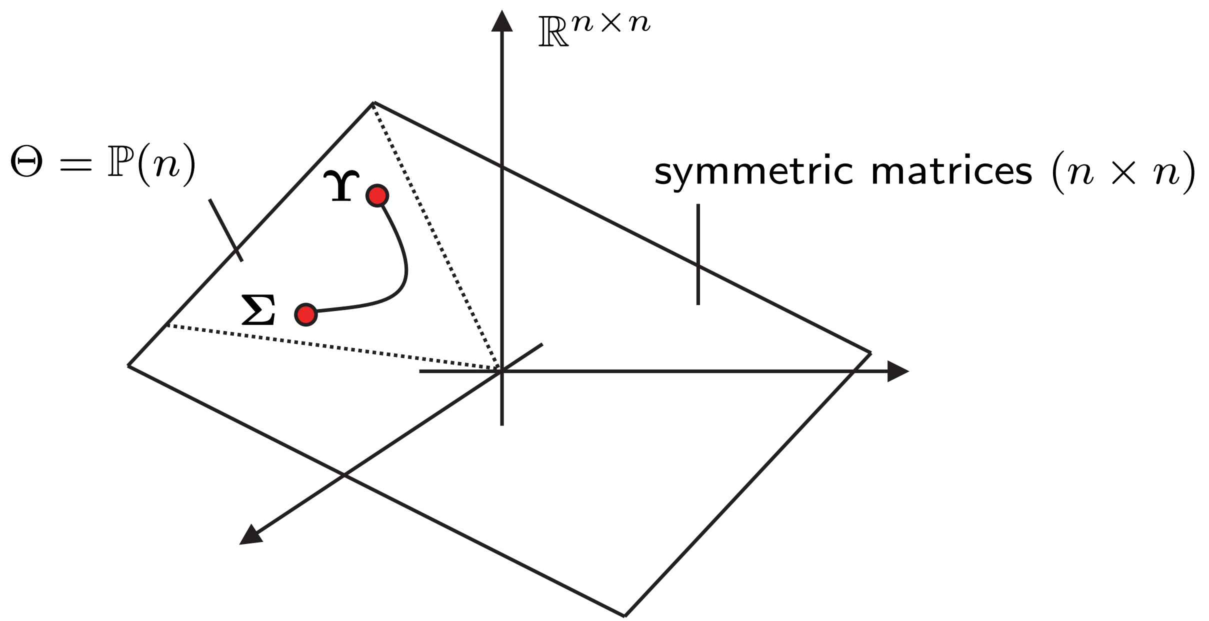


Information distance (geodesic)

Applications of DG: Information Geometry

$\diamond$ $\mathcal{F} = \{p(\boldsymbol{x}; \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, : \, \boldsymbol{\Sigma} \in \Theta = \mathbb{P}(n)\}$ ($\boldsymbol{\mu}$ is fixed)

$$d(\boldsymbol{\Sigma}, \boldsymbol{\Upsilon}) = \sqrt{\frac{1}{2}\sum_{i=1}^{n} (\log\, \lambda_i)^2}, \quad (\lambda_1, \ldots, \lambda_n) = \text{generalized eigenvalues of } (\boldsymbol{\Sigma}, \boldsymbol{\Upsilon})$$



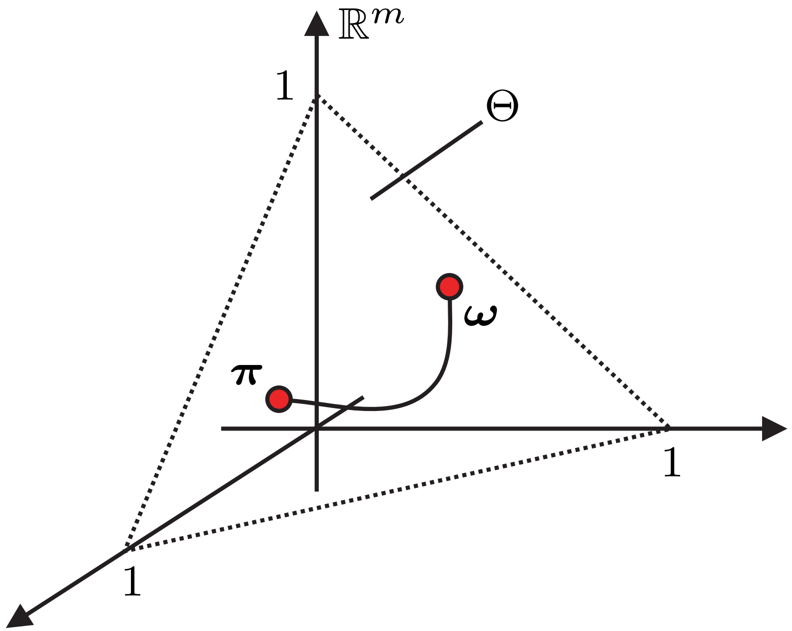
☐ Recall: $\mathbb{P}(n)$ = set of $n \times n$ positive definite matrices

Applications of DG: Information Geometry

- $\mathcal{F} = \{p(\boldsymbol{x};\boldsymbol{\pi}) \sim \text{multinomial}(n,\boldsymbol{\pi}) \;:\; \boldsymbol{\pi} \in \Theta = \text{simplex}(\mathbb{R}^m)\}$

  $\boldsymbol{x} = (x_1,\ldots,x_m) \in \mathbb{N}^m$, $\sum_{i=1}^m x_i = n$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_m)$, $\sum_{i=1}^m \pi_i = 1$

$$p(\boldsymbol{x};\boldsymbol{\pi}) = \frac{n!}{x_1!\cdots x_m!}\pi_1^{x_1}\cdots\pi_m^{x_m} \qquad d(\boldsymbol{\pi},\boldsymbol{\omega}) = 2\sqrt{n}\arccos\left(\sum_{i=1}^m \pi_i\omega_i\right)$$



Applications of DG: Information Geometry

- **Bibliography:**

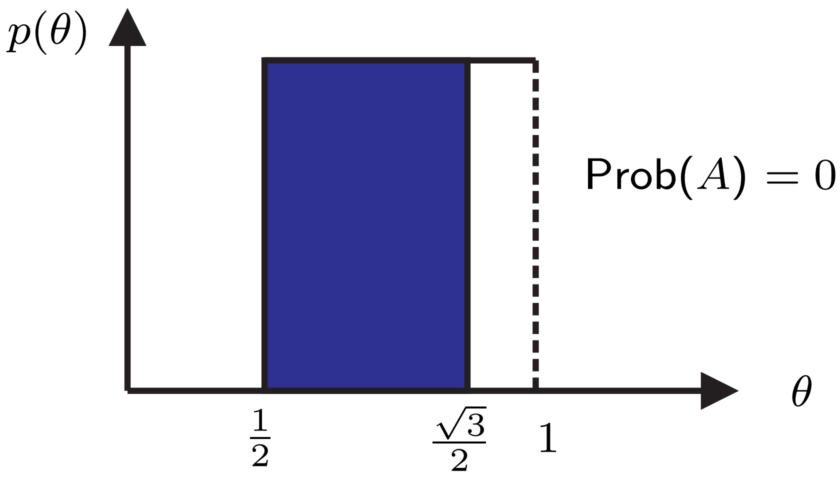
  - "Differential Geometry and Statistics", M. Murray *et al.*, Chapman & Hall, 1993
  - "The geometry of asymptotic inference", R. Kass, Statistical Science, vol. 4, no. 3, pp. 188–234, 1989
  - "Differential Geometry in Statistical Inference", S. Amari *et al.*, Institute of Mathematical Statistics, Lecture Notes, 1987
  - "The role of differential geometry in statistical theory", O.E. Barndorff-Nielsen *et al.*, International Statistical Review, 54, pp. 83–96, 1986
  - "Information and accuracy attainable in the estimation of statistical parameters", C. Rao, Bull. Calcutta Math. Soc., 37, pp. 81–91, 1945

Applications of DG: Geometrical Interpretation of Jeffreys' Prior

☐ **Problem:** Given a parametric statistical family $\mathcal{F} = \{p(\boldsymbol{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ assign a non-informative prior $p(\boldsymbol{\theta})$ for the parameter $\boldsymbol{\theta}$

☐ **Example:** $\mathcal{F} = \left\{p(x;\theta) \sim \mathcal{N}(0,\theta^2) : \theta \in \Theta = (1/2, 1)\right\}$

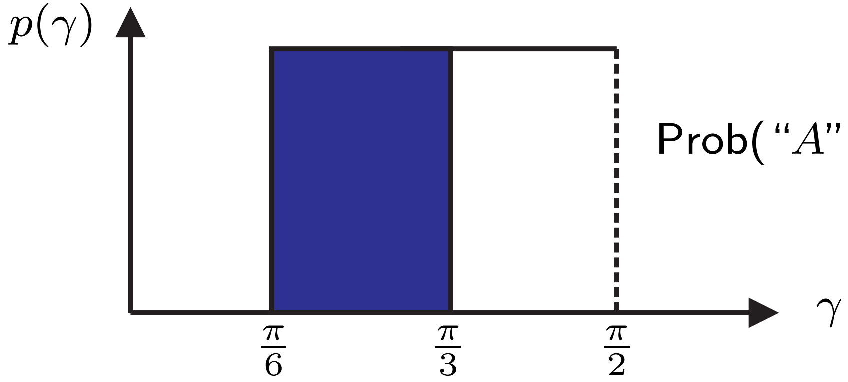
Naive choice (uniform distribution):



☐ This method does not produce "intrinsic" priors (parameter invariant)

Applications of DG: Geometrical Interpretation of Jeffreys' Prior

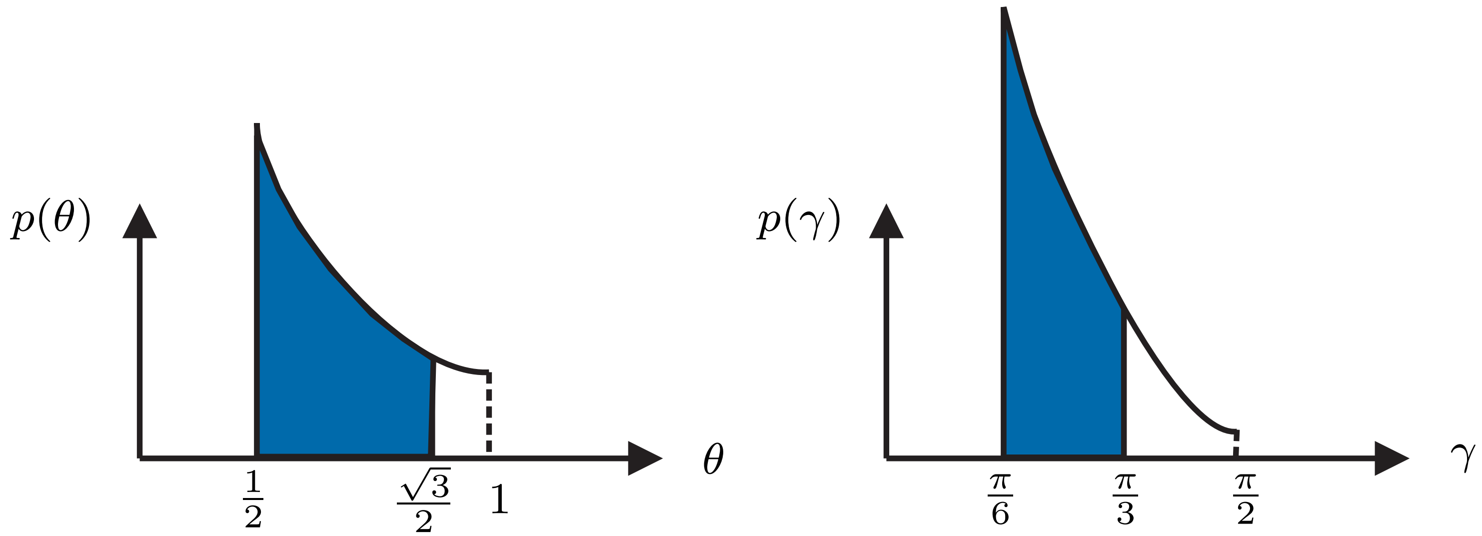
☐ With $\theta = \sin(\gamma)$: $\mathcal{F} = \left\{p(x;\gamma) \sim \mathcal{N}(0, \sin^2(\gamma)) \ : \ \gamma \in \Gamma = (\pi/6, \pi/2)\right\}$



☐ Jeffreys' prior: $p(\boldsymbol{\theta}) \propto \sqrt{\det(\boldsymbol{I}(\boldsymbol{\theta}))}$ where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information matrix

Applications of DG: Geometrical Interpretation of Jeffreys' Prior

- **For the current example:** $p(\theta) \propto \frac{1}{\theta}$ and $p(\gamma) \propto \text{cotg}(\gamma)$

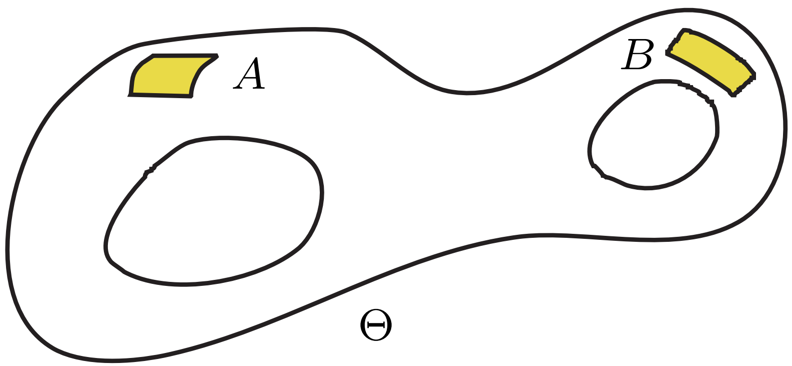


$$\text{Prob}(A) = \text{Prob}(\text{“}A\text{”}) = 0.79$$

Applications of DG: Geometrical Interpretation of Jeffreys' Prior

- **Differential geometric interpretation:** Jeffreys' prior is simply the Riemannian volume element induced by the Fisher metric!

- **Insight:** A parametric statistical family is an autonomous geometrical object carrying its own "uniform" prior (applies equal mass to sets of equal area)



$$\text{Area}(A) = \text{Area}(B) \Rightarrow \text{Prob}(\boldsymbol{\theta} \in A) = \text{Prob}(\boldsymbol{\theta} \in B)$$

Applications of DG: Geometrical Interpretation of Jeffreys' Prior

□ **Bibliography:**

◇ "The geometry of asymptotic inference", R. Kass, Statistical Science, vol. 4, no. 3, pp. 188–234, 1989

◇ "Differential Geometry in Statistical Inference", S. Amari *et al.*, Institute of Mathematical Statistics, Lecture Notes, 1987

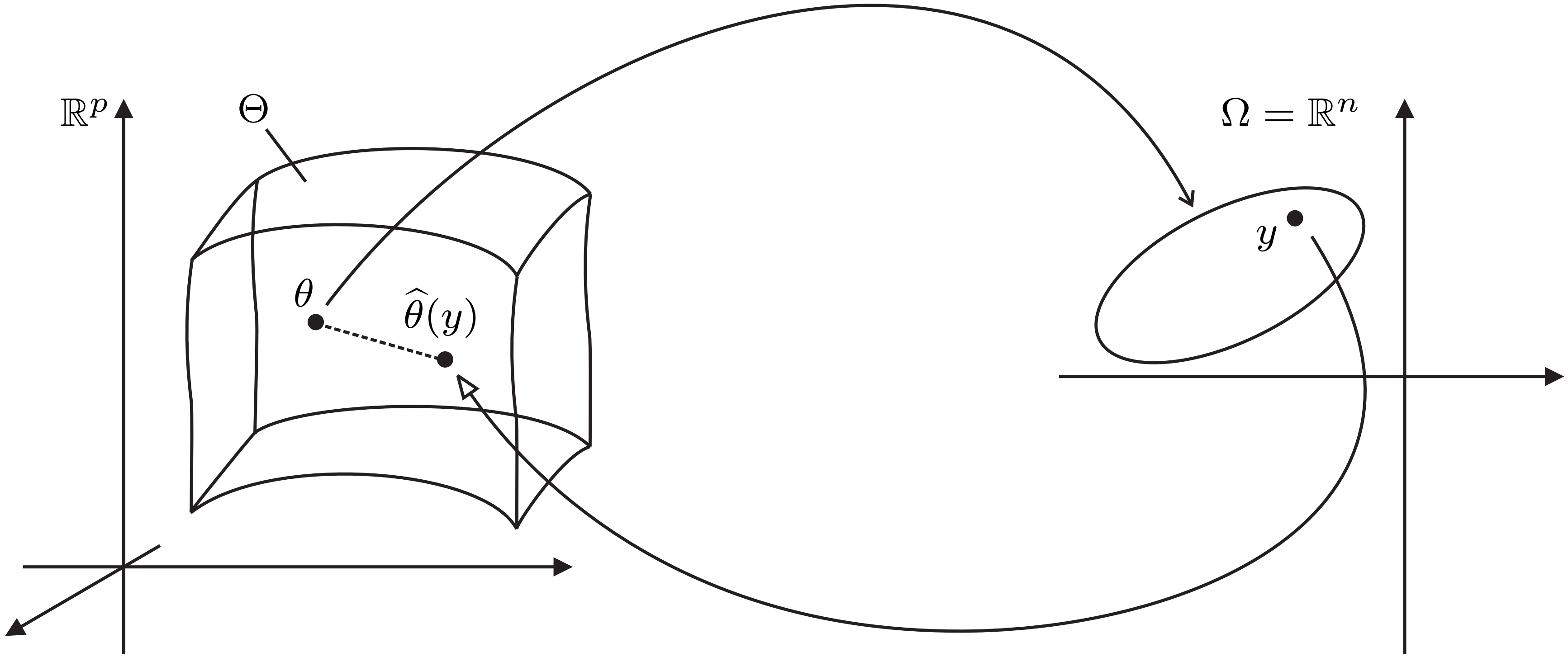
◇ "The role of differential geometry in statistical theory", O.E. Barndorff-Nielsen *et al.*, International Statistical Review, 54, pp. 83–96, 1986

◇ "Theory of Probability", 3rd ed., H. Jeffreys, Oxford University, 1961

◇ "An invariant form for the prior probability in estimation problems", H. Jeffreys, Proc. Royal Soc. London Ser. A, 196, pp. 453–461, 1946

Application of DG: bounds

☐ **Classical Euclidean setup:**

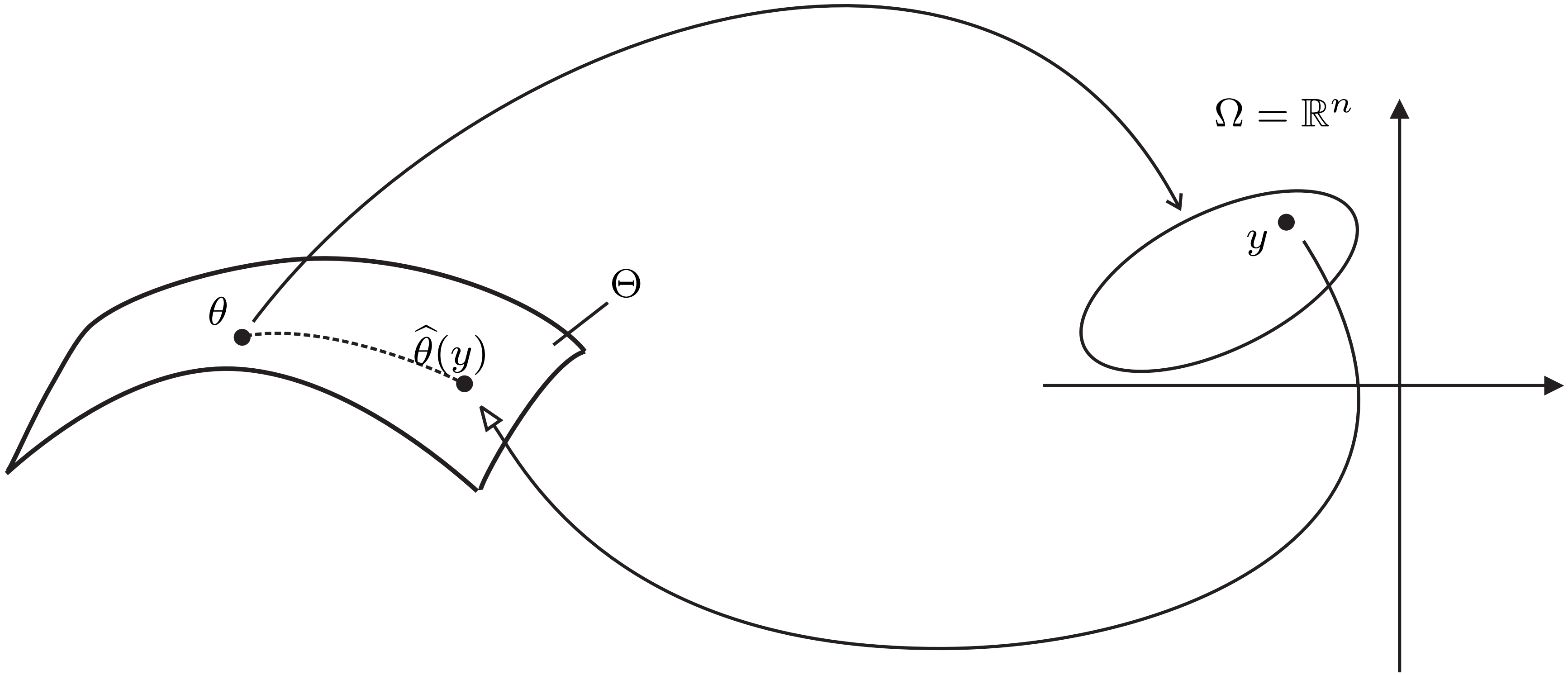


☐ **Cramér-Rao Bound (CRB):**

$$\mathsf{var}_\theta\left(\widehat{\theta}\right) = \mathsf{E}_\theta\left\{d\left(\theta, \widehat{\theta}(Y)\right)^2\right\} \geq \mathsf{tr}\left(I_\theta^{-1}\right) \qquad (I_\theta = \text{ Fisher matrix })$$

Application of DG: bounds

☐ **Riemannian setup:**



☐ **Intrinsic Variance Lower Bound (IVLB):**

$$\mathsf{var}_\theta\left(\widehat{\theta}\right) = \mathsf{E}_\theta\left\{d\left(\theta, \widehat{\theta}(Y)\right)^2\right\} \geq ?$$

Applications of DG: bounds

☐ **Theorem (IVLB).** Suppose:

▷ The sectional curvature of Θ is upper bounded by $C \geq 0$

▷ $+$ some technical conditions

Then,

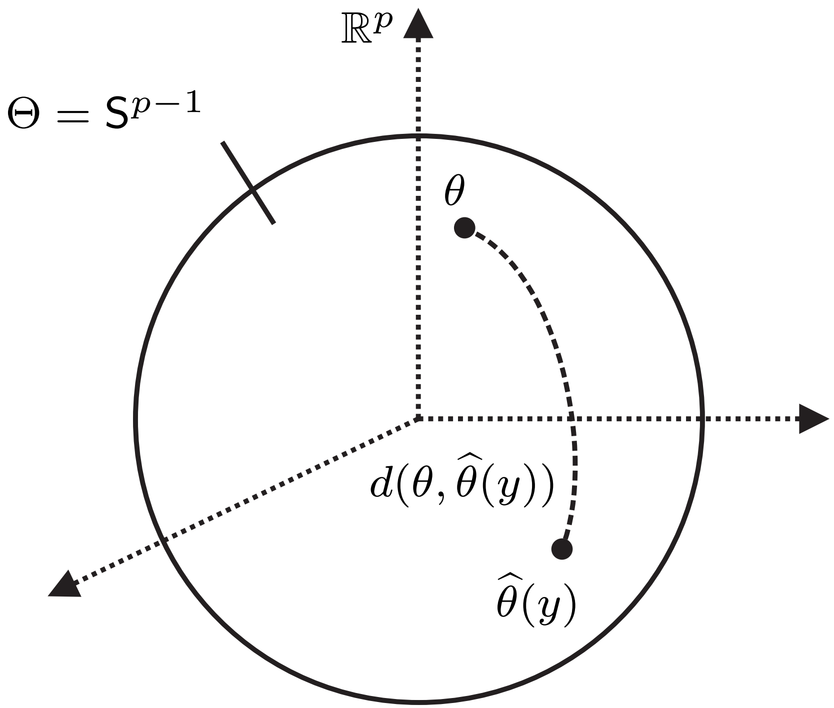
$$\mathrm{var}_\theta\left(\widehat{\theta}\right) \geq \begin{cases} \lambda_\theta & , \quad \text{if } C = 0 \\ \dfrac{\lambda_\theta C + 1 - \sqrt{2\lambda_\theta C + 1}}{C^2 \lambda_\theta / 2} & , \quad \text{if } C > 0 \end{cases}$$

where:

▷ $\lambda_\theta = \mathrm{tr}(I_\theta^{-1}) \qquad (I_\theta = \text{ Fisher tensor })$

Example: inference on S^{p-1}

☐ $\mathsf{S}^{p-1} = \{x \in \mathbb{R}^p \,:\, \|x\| = 1\}$ is the unit-sphere in $\mathbb{R}^p$



☐ **Geometry of** Θ: $d(\theta, \widehat{\theta}(y)) = \mathrm{acos}(\theta^T \widehat{\theta}(y))$ and $C = 1$

Example: inference on S^{p-1}

☐ **Observation**: $y = \theta + w \in \mathbb{R}^p \ (p = 10)$

▷ $\theta \in \Theta = \mathsf{S}^{p-1}$

▷ $w \sim \mathcal{N}(0, \sigma^2 I_p)$

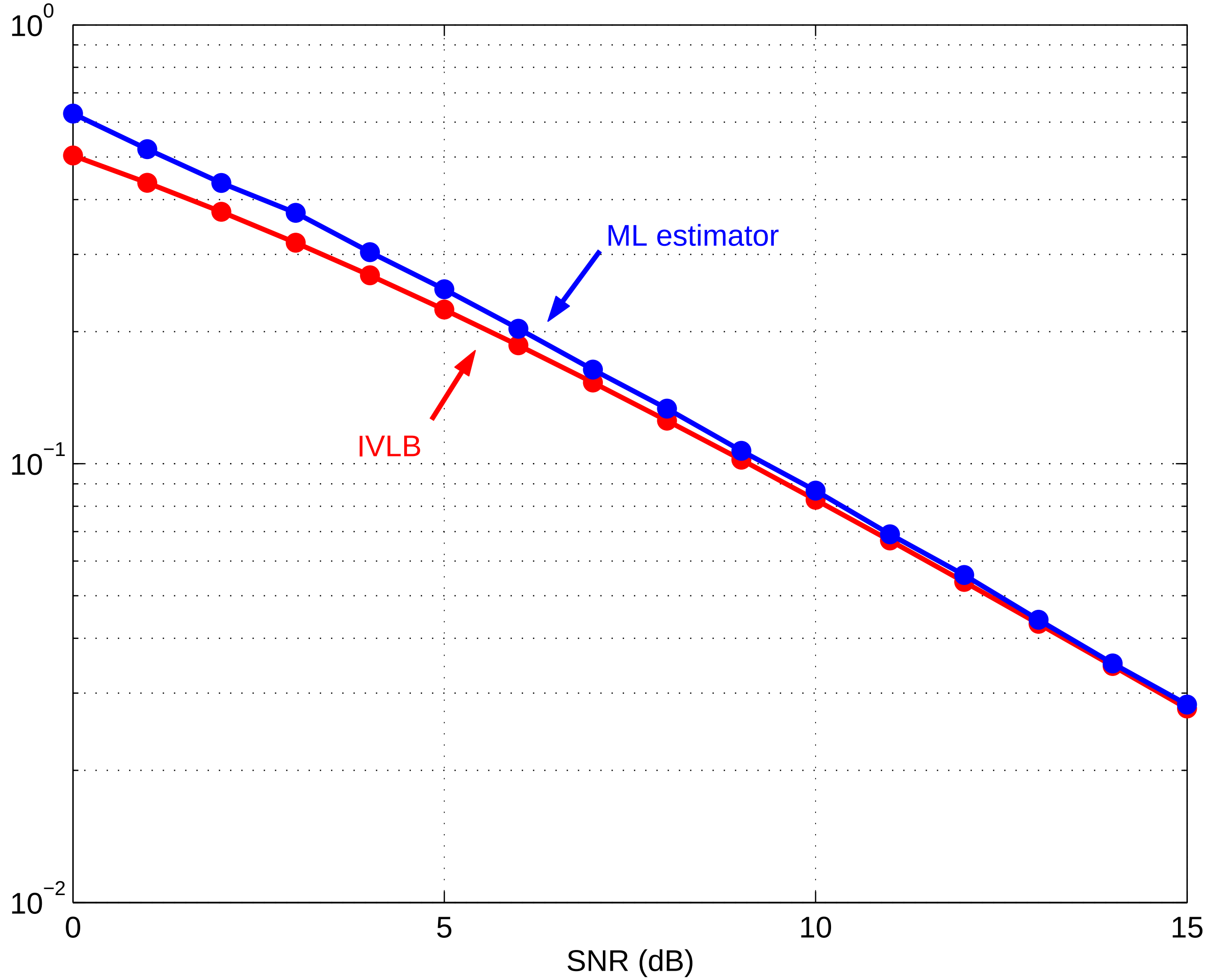
☐ Maximum-likelihood estimator:

$$\widehat{\theta}(y) = \frac{y}{\|y\|}$$

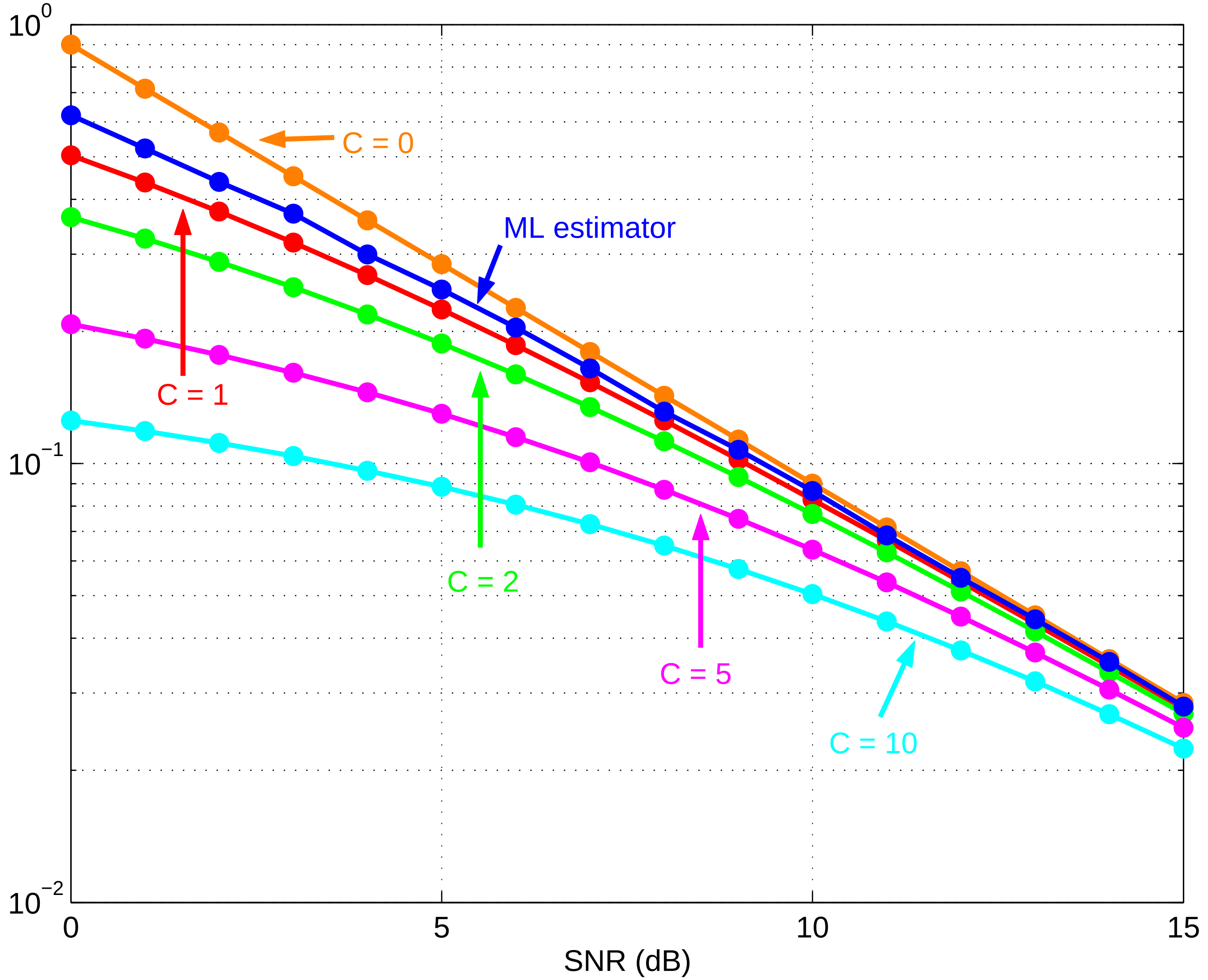
☐ Signal-to-noise ratio:

$$\mathsf{SNR} = \frac{\mathsf{E}\left\{\|\theta\|^2\right\}}{\mathsf{E}\left\{\|w\|^2\right\}} = \frac{1}{p\,\sigma^2}$$

Example: inference on $\mathbb{S}^{p-1}$



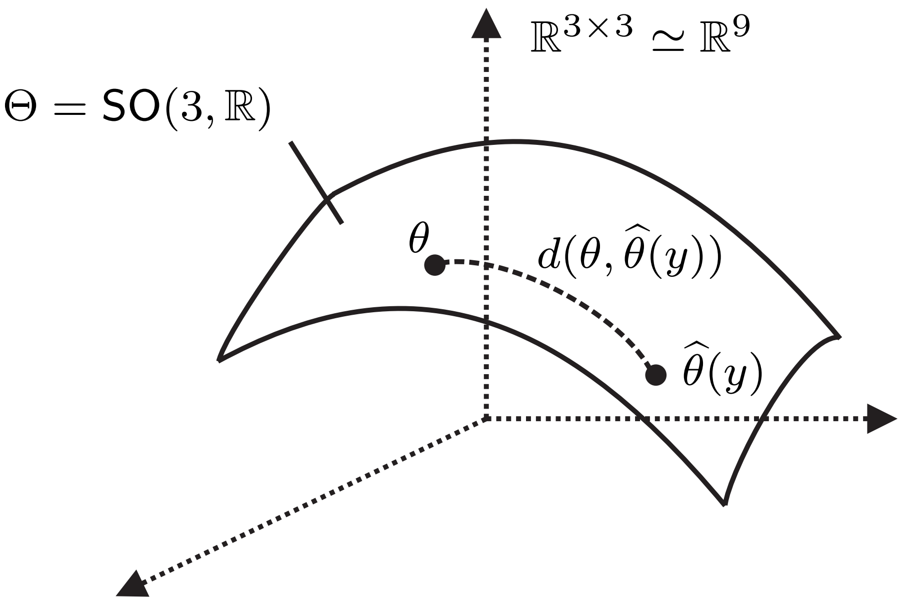
Example: inference on $\mathbb{S}^{p-1}$



Example: inference on $\mathsf{SO}(3,\mathbb{R})$

- $\mathsf{SO}(3,\mathbb{R})$ is the special orthogonal group:

$$\mathsf{SO}(3,\mathbb{R}) = \left\{ Q \in \mathbb{R}^{3\times 3} \,:\, Q^T Q = I_3, \det(Q) = 1 \right\}$$



- **Geometry of** Θ: $d(\theta, \widehat{\theta}(y)) = \sqrt{2}\,\mathsf{acos}(0.5[\mathsf{tr}(\theta^T \widehat{\theta}(y)) - 1])$ and $C = 1/8$

Example: inference on $\mathsf{SO}(3, \mathbb{R})$

- **Observation**: $Y = \theta X + W \in \mathbb{R}^{3 \times k}$ $(k = 10)$
  - ▷ $\theta \in \Theta = \mathsf{SO}(3, \mathbb{R})$: unknown rotation matrix [Procrustean analysis]
  - ▷ $X = [\, x_1 \, x_2 \, \cdots \, x_k \,]$: constellation of known k landmarks in $\mathbb{R}^3$ ($X X^T = I_3$)
  - ▷ $W = [\, w_1 \, w_2 \, \cdots \, w_k \,]$, $w_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_3)$: additive observation noise

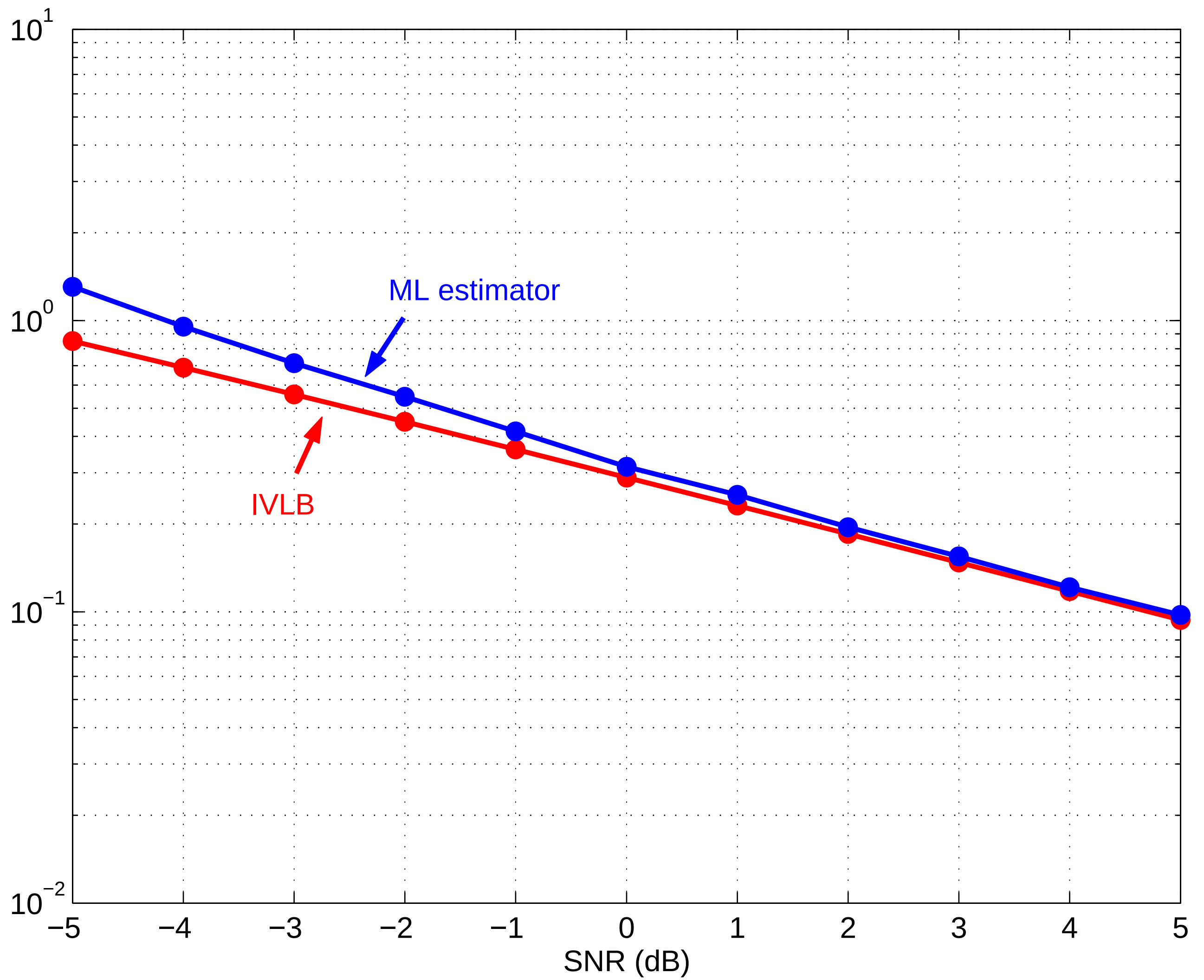
- Maximum-likelihood estimator:

$$\widehat{\theta}(Y) = \cdots \; (\text{closed} - \text{form})$$

- Signal-to-noise ratio:

$$\mathsf{SNR} = \frac{\mathsf{E}\left\{ \|\theta X\|^2 \right\}}{\mathsf{E}\left\{ \|W\|^2 \right\}} = \frac{1}{k\,\sigma^2}$$

Example: inference on $\mathrm{SO}(3,\mathbb{R})$



Applications of DG: Bounds

Bibliography:

- "Covariance, subspace, and intrinsic Cramér-Rao bounds," S. Smith, IEEE Trans. on Signal Proc., vol. 53, no.5, May 2005
- "Intrinsic variance lower bound (IVLB): an extension of the Cramér-Rao bound to Riemannian manifolds", J. Xavier and V. Barroso, IEEE Int. Conf. on Acoust., Sp. and Sig. Proc. (ICASSP), March 2005
- "The Riemannian geometry of certain parameter estimation problems with singular Fisher matrices", J. Xavier and V. Barroso, IEEE Int. Conf. on Acoust., Sp. and Sig. Proc. (ICASSP), May 2004
- "Hilbert-Schmidt lower bounds for estimators on matrix Lie groups for ATR", U. Grenander *et al.*, IEEE Trans. on Patt. Anal. and Mach. Intell., vol. 20, no. 8, pp. 790–801, August 1998
- "On the Cramér-Rao bound under parametric constraints", P. Stoica *et al.*, IEEE Sig. Proc. Lett., vol. 5, no. 7, pp. 177–179, July 1998
- "Intrinsic analysis of statistical estimation", J. Oller *et al.*, The Annals of Stat., vol. 23, no. 5, pp. 1562–1581, 1995
- "A Cramér-Rao type lower bound for estimators with values in a manifold", H. Hendricks, Journal of Multivar. Anal., no. 38, pp. 245–261, 1991

Course's Table of Contents

☐ **Three main topics:**

- ▷ Topological manifolds
- ▷ Differentiable manifolds
- ▷ Riemannian manifolds

☐ **Three layers of structure:**

Riemannian structure	Length of curves ; Geodesics ; Distance ; Connections ; etc
Differentiable structure	Tangent vectors; Smooth maps; Tensors; Integration ; etc
Topological structure	Boundary of sets; Convergent sequences; Continuous maps ; etc
Plain set	

Course's Table of Contents

☐ **Topological manifolds:** "Introduction to Topological Manifolds", J. Lee, Springer-Verlag

- ◇ Ch.2: Topological spaces
- ◇ Ch.3: New spaces from old
- ◇ Ch.4: Connectedness and compacteness

☐ **Smooth manifolds:** "Introduction to Smooth Manifolds", J. Lee, Springer-Verlag

- ◇ Ch.2: Smooth maps
- ◇ Ch.3: The tangent bundle
- ◇ Ch.5: Submanifolds
- ◇ Ch.7: Lie group actions
- ◇ Ch.8: Tensors
- ◇ Ch.9: Differental forms
- ◇ Ch.10: Integration on manifolds

Course's Table of Contents

Bibliography for the Course

☐ **Topological manifolds**

- ◇ “Introduction to Topological Manifolds”, J. Lee, Springer-Verlag, 2000
- ◇ “Introduction to Topology and Modern Analysis”, G. Simmons, 1963

☐ **Smooth manifolds**

- ◇ “Introduction to Smooth Manifolds”, J. Lee, Springer-Verlag, 2002
- ◇ “ An Introduction to Differentiable Manifolds and Riemannian Geometry”, 2nd ed., W.Boothby, Academic Press, 1986
- ◇ “Manifolds, Tensor Analysis and Applications”, R. Abraham *et al.*, Springer-Verlag, 1988
- ◇ “A Comprehensive Introduction to Differential Geometry”, vol.I, M. Spivak, Publish or Perish, 1979
- ◇ “Lectures on Differential Geometry”, S. Chern, W. Chern and K. Lam, World Scientific, 1999

☐ **Riemannian manifolds**

- ◇ “Riemannian Manifolds”, J. Lee, Springer-Verlag
- ◇ “Riemannian Geometry”, M. Carmo, Birkhauser, 1992

Bibliography

□ **Other references (introductory):**

- "Differential Forms with Applications to the Physical Sciences", H. Flanders, Dover, 1963
- "Differential Forms with Applications", M. Carmo, Springer-Verlag, 1994

□ **Other references (advanced):**

- "Riemannian Geometry", S. Gallot, D. Hulin and J. Lafontaine, Springer-Verlag, 1987
- "A Comprehensive Introduction to DG", vol.II-V, M. Spivak, Publish or Perish, 1979
- "Riemannian Geometry: A Modern Introduction", I. Chavel, Cambridge Press, 1993
- "Riemannian Geometry and Geometric Analysis", J. Jost, Springer-Verlag, 1998
- "Foundations of Differential Geometry", vol. I-II, S. Kobayashi and K. Nomizu, Wiley 1969
- "DG, Lie Groups and Symmetric Spaces", S. Helgason, Academic Press, 1978

□ **Many others...**

Grading

☐ **Grade** = Homework (60%) + Project (40%)

☐ **Homeworks:**

#	Received	Due
1	March, 29	April, 19
2	April, 19	May, 10
3	May, 10	May, 31
4	May, 31	June, 21

☐ **Project (individual):** A paper will be assigned for each student to study

Output: public presentation of the paper

Start: May, 10 End: July, 31

$\boxed{\textbf{Discussion, questions, etc}}$