Une introduction à la géométrie de l'information

Frank Nielsen

23 février 2022

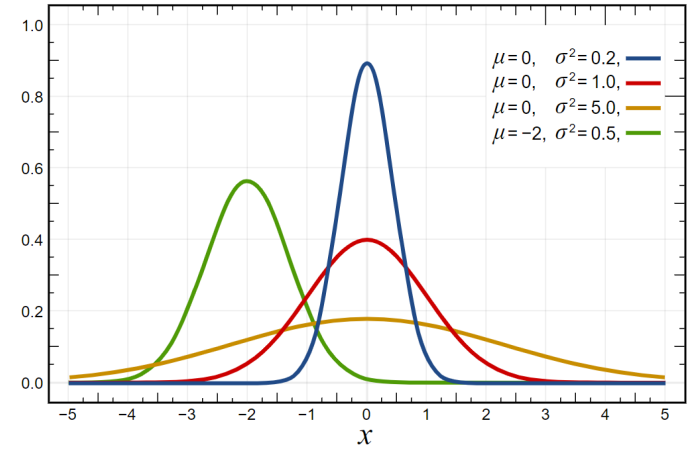
Sony CSL

COLLÈGE DE FRANCE 1530

Qu'est-ce que la géométrie de l'information ? (1/4)

Considérons une famille de lois de distributions : le **modèle statistique**

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \lambda = (\mu, \sigma) \in \mathbb{H} \right\}$$

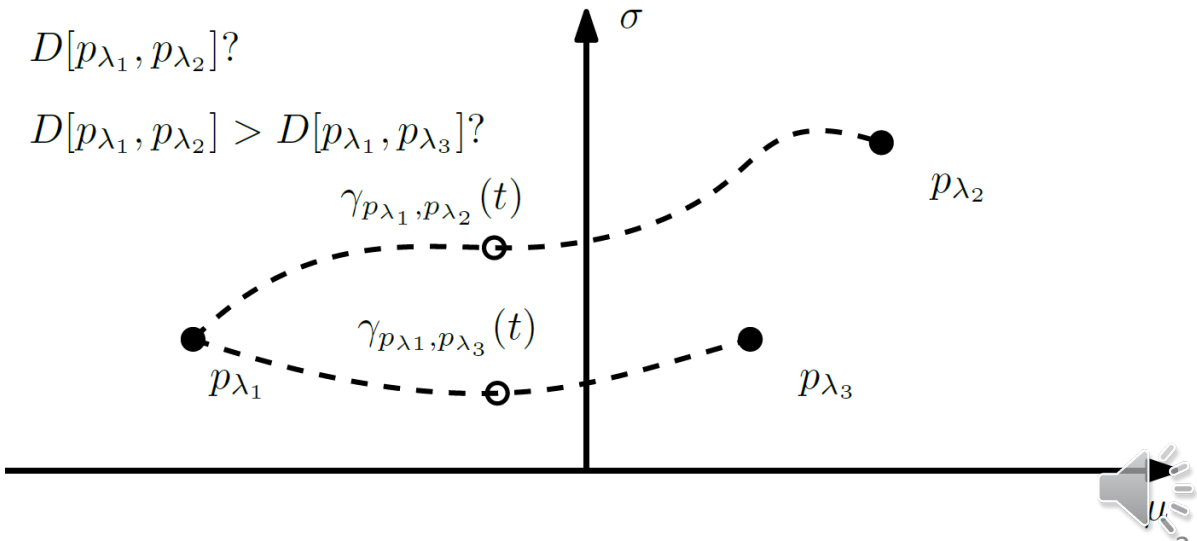


Espace des paramètres Λ : le demi plan $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$

Quelles sont les **structures géométriques** pour cette famille de lois normales ?

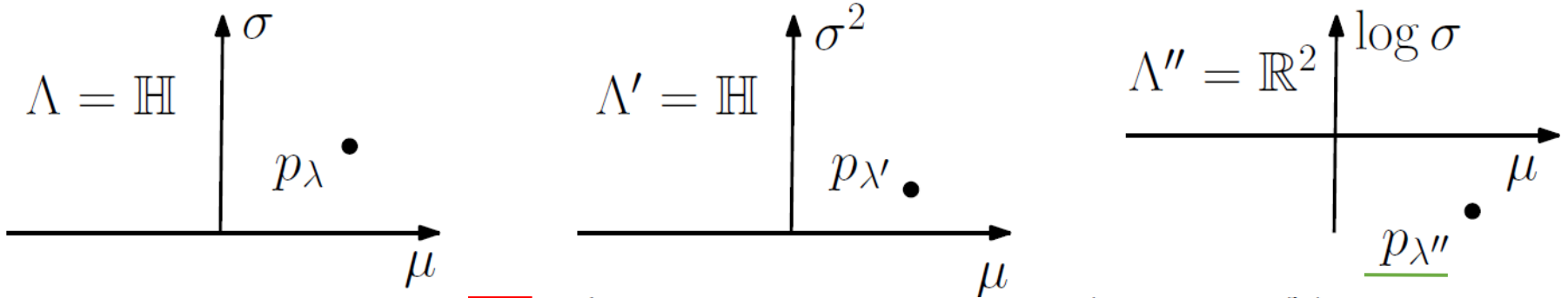
Quelques questions:

- Comment **interpoler** entre deux lois ?
- Quelle **distance** entre deux lois ?
- Y a-t-il **plusieurs façons** de faire ?
  et si oui, pourquoi ?



Qu'est-ce que la géométrie de l'information ? (2/4)

- Quelles **invariances** doivent satisfaire les **structures géométriques** et les **distances** entre lois ?
- **Premier principe d'invariance** : Si on indexe les lois normales par **(μ,σ^2)** ou **$(\mu,\log(\sigma))$** au lieu de **(μ,σ)**, cela ne doit pas changer ni les distances ni les chemins d'interpolation appelés ``**géodésiques**''



Même famille de lois :

$$\mathcal{P} = \left\{ p_{\lambda''}(x) = \frac{1}{\sqrt{2\pi}\exp(\lambda_2'')} \exp\left(-\frac{(x-\lambda_1'')^2}{2\exp(2\lambda_2'')}\right), \quad \lambda'' = (\mu, \log\sigma) \in \mathbb{R}^2 \right\}$$

On désire donc avoir ces **deux invariances** :

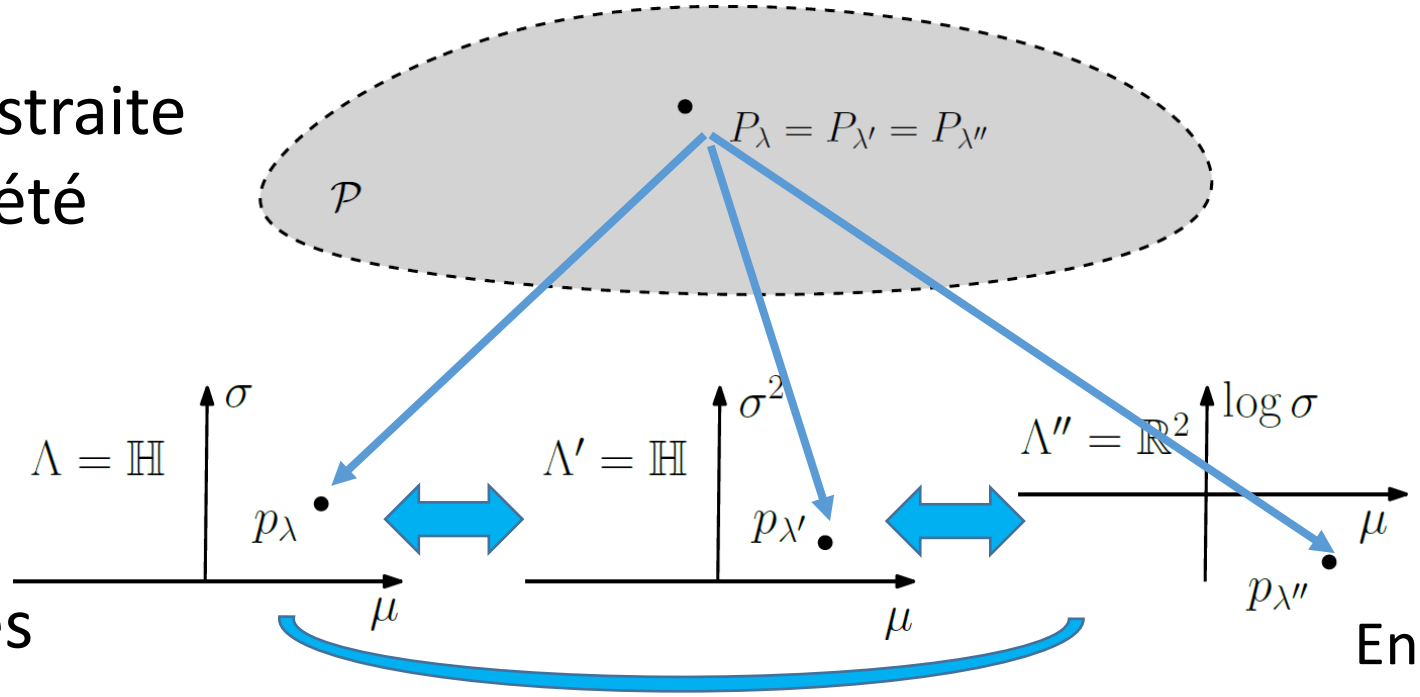
① $D[p_{\lambda_1}, p_{\lambda_2}] = D[p_{\lambda_1'}, p_{\lambda_2'}] = D[p_{\lambda_1''}, p_{\lambda_2''}]$

② $\gamma_{p_{\lambda_1}, p_{\lambda_2}}(t) = \gamma_{p_{\lambda_1'}, p_{\lambda_2'}}(t) = \gamma_{p_{\lambda_1''}, p_{\lambda_2''}}(t), \forall t \in [0,1]$

Variétés différentielles des modèles statistiques

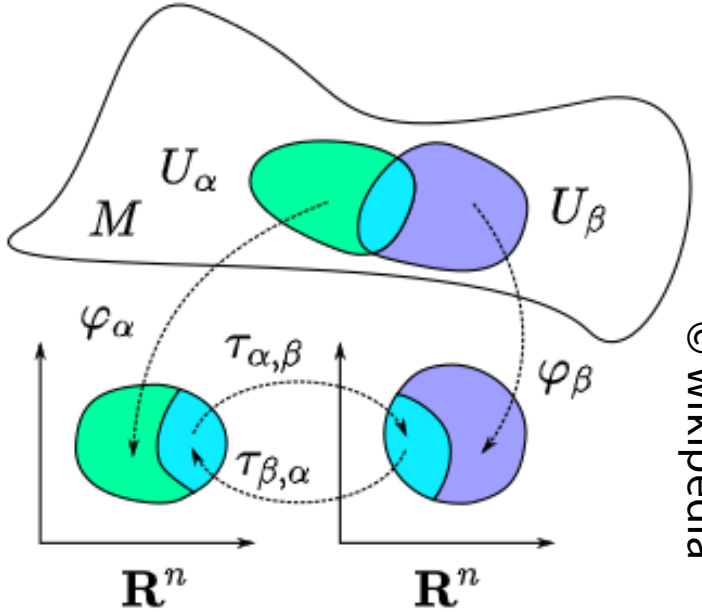
- À chaque point de la variété correspond une unique loi paramétrique
- Le modèle est dit **identifiable** lorsque $\lambda \leftrightarrow P_\lambda$
- Souvent une seule **carte** globale = atlas = domaine des paramètres

Figure abstraite de la variété



Domaines

Différentes cartes globales (atlas à une carte)



© wikipedia

En général, plusieurs cartes pour naviguer sur les variétés (ex., deux cartes pour la sphère)

Qu’est-ce que la géométrie de l’information ? (3/4)

- La **géométrie de l'information** : étude de structures géométriques sur les variétés induites par les modèles statistiques
- utilisation du jargon géométrique comme les **géodésiques**, les **boules**, la **projection informatielle** ou la **courbure statistique** en utilisant le **calcul tensoriel** (qui a permis d’étudier l’efficacité d’estimateurs en statistique aux ordres supérieurs)
- Étude des principes de l'**invariance** en statistique
- Les nouvelles structures géométriques duales peuvent s'appliquer **en dehors du cadre statistique** aussi, par exemple en optimisation sur les variétés

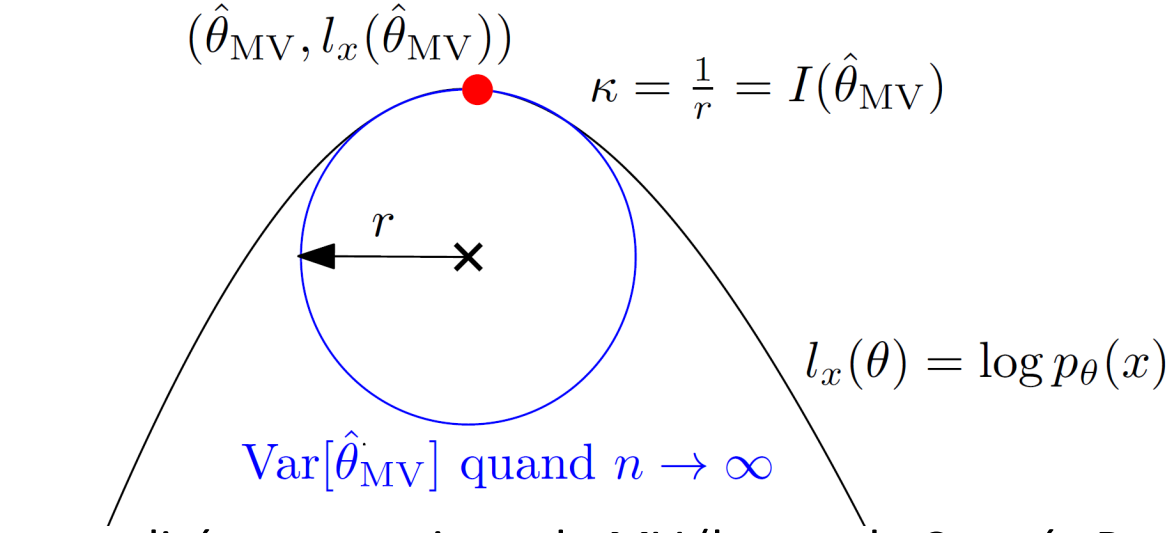
- La géométrie de l'information est née de la curiosité d'équiper une variété Riemannienne de la **métrique de Fisher**, et d'utiliser la **distance géodésique** pour résoudre des problèmes de classification ou de tests d'hypothèses en statistique **[Mahalanobis 1936] [Hotelling 1930] [Rao 1945]**

L'information de Fisher $I_X(\theta)$

- Pour une famille paramétrique de lois à D paramètres, la **matrice d'information de Fisher** (MIF) est définie comme la matrice de covariance du **score**

$$X \sim p_\theta(x) \qquad s_X(\theta) = \nabla_\theta \log p_\theta(x) \qquad \boxed{I_X(\theta) = \mathrm{Cov}[s_\theta]}$$

- La matrice de Fisher est *symmétrique semi-définie positive*
- La MIF est dite **régulière** quand elle est finie et définie-positive
- Interprétée comme la **courbure** du graphe de la fonction log-vraisemblance:



normalité asymptotique du MV (borne de Cramér-Rao)

En général, pour une fonction arbitraire f(x), la courbure est :

$$\kappa(x) = \frac{f''(x)}{(1 + (f'(x))^2)^{\frac{3}{2}}}, \qquad r(x) = \frac{(1 + (f'(x))^2)^{\frac{3}{2}}}{|f''(x)|}$$

MV: Maximum de Vraisemblance

Le **cercle osculateur** à son rayon inversement proportionel à la courbure

Les deux formes usuelles de la matrice de Fisher

- En utilisant les deux premières **identité de Bartlett** sous les conditions de régularités d'échange k-fois des opérations de différentiation avec l'intégration

(Bartlett k) $$\nabla^k E_\theta[\exp(l_x(\theta))] = E_\theta\left[\nabla^k \exp(l_x(\theta))\right] = \nabla^k E_\theta[p_\theta] = \nabla^k 1 = 0$$

- Ré-écrit la matrice d'information de Fisher sous ses formes les plus connues :

$I_X(\theta) = \mathrm{Cov}[s_\theta]$ $\quad \mathrm{Cov}[s_\theta] = E[s_\theta s_\theta^\top] - E[s_\theta]E[s_\theta]^\top$

① **Première forme** : $$E[\nabla l_x(\theta)] = 0 \Rightarrow I_X(\theta) = E_\theta\left[\nabla \log p_\theta(x)(\nabla \log p_\theta(x))^\top\right]$$

(Bartlett k=1)

② **Deuxième forme** (moins la Hessienne de la log-vraisemblance) :

$$E_\theta\left[\nabla \log p_\theta(x)(\nabla \log p_\theta(x))^\top\right] + E\left[\nabla^2 \log p_\theta(x)\right] = 0 \Rightarrow I_X(\theta) = -E[\nabla^2 \log p_\theta(x)]$$

(Bartlett k=2)

$$I_X(\theta) = -E_{p_\theta}\left[\nabla^2 l_x(\theta)\right] = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} l_x(\theta) l_x(\theta)\right]_{ij}$$

Mahalanobis et la ``distance généralisée''

© wikipedia

P. C. Mahalanobis
(1893-1972)
Fondateur de
l'Institut de Statistique Indien (ISI)

- Analyse de données en craniologie sur des groupes : chaque crâne est caracterisé par d attributs.

- Mahalanobis (1928, 1936) propose d'utiliser cette **divergence** entre deux groupes S_1 et S_2 :

$$\Delta^2[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] = (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)$$

Matrice de précision

- Aujourd'hui, la **distance de Mahalanobis** (métrique) :

$$\Delta[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)}$$

qui généralise la **distance euclidienne** quand Σ =I, matrice identité

- Divergence = dissimilarité lisse qui ne satisfait pas **l'inégalité triangulaire** des distances métriques

Vol. VIII. } APRIL & SEPT. 1928. { Nos. 2 & 3.

I.—A STATISTICAL STUDY OF THE CHINESE HEAD

BY

P. C. MAHALANOBIS

ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

Les distances de Mahalanobis : Des espaces vectoriels équipés d'un produit scalaire

- La distance de Mahalanobis se ré-écrit comme $\Delta_\Sigma(\mu, \mu') = \|\mu - \mu'\|_{\Sigma^{-1}}$

$$\|x\|_{\Sigma^{-1}} = \sqrt{x^\top \Sigma^{-1} x}$$

- Pour une matrice symmétrique positive-definite Q, on définit le **produit scalaire** par la forme bilinéaire : $\langle v_1, v_2 \rangle_Q = v_1^\top Q v_2$

- Le produit scalaire induit une **norme** qui induit une **distance** métrique :

$$\langle v_1, v_2 \rangle_E \rightarrow \|v\|_E = \sqrt{\langle v, v \rangle} \rightarrow D_E(v_1, v_2) = \|v_1 - v_2\|_E$$

- Le produit scalaire permet de définir la notion d'**orthogonalité** entre deux vecteurs (plus généralement l'angle formé entre eux) ainsi que les **longeurs** de vecteurs : $v_1 \perp v_2 \leftrightarrow \langle v_1, v_2 \rangle_Q = 0$ $\|v\| = \sqrt{\langle v, v \rangle}$
- On va voir que c'est la géométrie des espaces tangents de la variété statistique

La métrique Riemannienne de Fisher

- Sur la variété $\mathcal{P} = \{p_\theta(x) \ : \theta \in \Theta\}$

g_F : champ de produits scalaires lisse des plans tangents

$$g_F(u,v) = [u]_B^\top \, I(\theta) \, [v]_B$$

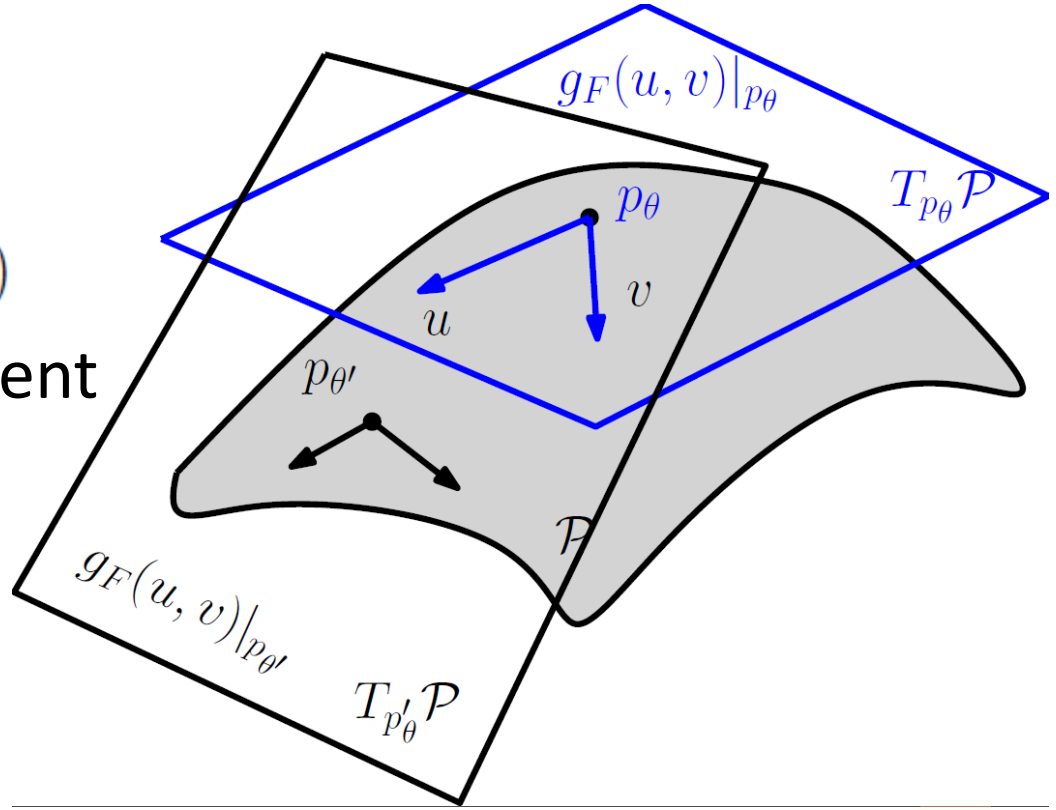
- **Composantes** du vecteur $[v]_B = (v^1, \ldots, v^D)$

exprimée dans la base naturelle du plan tangent induit par la carte θ(.) $\partial_i = \dfrac{\partial}{\partial \theta^i}$

$$B = \{e_1 = \partial_1, \ldots, e_D = \partial_D\}$$

$$g_{ij} = g_F(\partial_i, \partial_j) = I_{ij}(\theta)$$

$$g_F(u,v) = \sum_{i,j} g_{ij} u^i v^j = u^\top I(\theta) \, v$$



Représentation du plan tangent pour une variété issue d'un modèle statistique

- Dans le plan tangent, on peut choisir parmi une infinité de bases !
- La métrique est invariante par changement de base : seules les **composantes** des vecteurs changent.
- On exprime un vecteur v par une **représentation** v(x)
- Les vecteurs de la base T_θ sont les **vecteurs scores** :

$$B = \{e_1 = \partial_1 l_x(\theta), \ldots, e_D = \partial_D l_x(\theta)\}$$

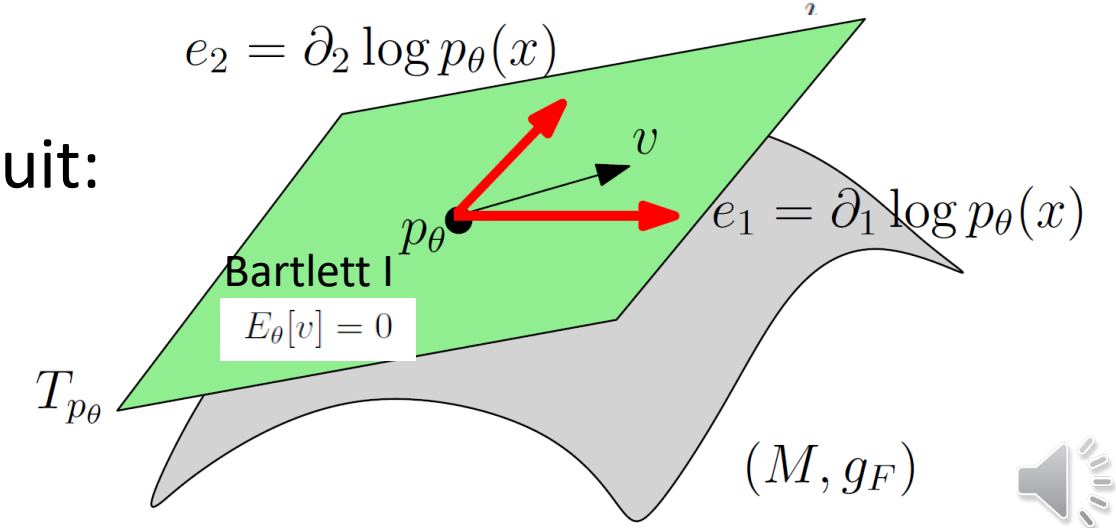
- Le produit scalaire se réinterprète comme suit:

$$g_F(u,v) = E_\theta[u(x)v(x)] = \mathrm{Cov}(u(x), v(x))$$

$$g_F(\partial_i, \partial_j) = E_\theta[\partial_i l_x(\theta)\partial_j l_x(\theta)]$$

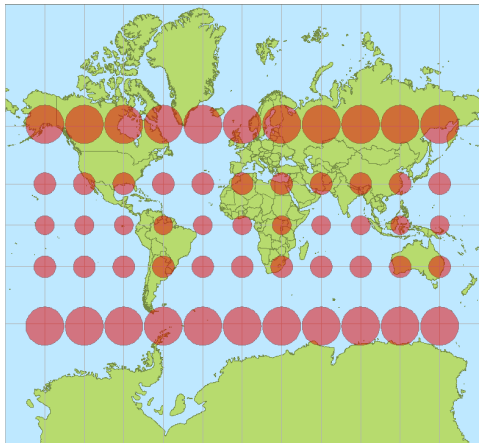
Espérance

$$T_\theta = T_{p_\theta} = \{\sum_i v^i \partial_i l_x(\theta)\}$$



Visualiser la métrique de Fisher et la borne de Cramér–Rao

- Métrique de Fisher : $g_F(u,v) = [u]_B^\top I(\theta)\,[v]_B$
- On visualise I(θ) par un **ellipsoide**
- On visualise le champ de tenseur métrique avec les **indicatrices de Tissot** :



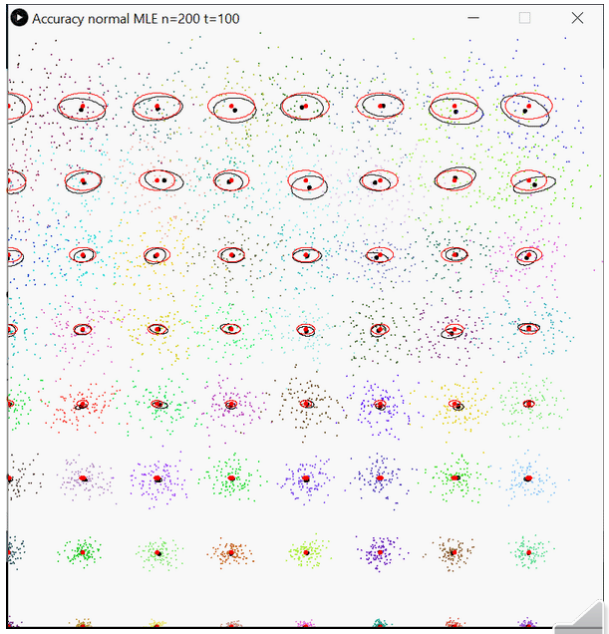
Visualiser la borne de CR : $\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta)^{-1}$

Inverse de la MIF (indicatrices de Tissot)

- Pour chaque position d'une grille (μ,σ) :

- génére iid échantillons iid N (μ,σ)

$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$



- estime le maximum de vraisemblance

- répète k fois pour obtenir ***un estimateur empirique de la matrice de covariance***

$$\mathrm{Cov}[\hat{\theta}_n]$$

demi plan (μ,σ) $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$

La distance de Rao sur la variété de Fisher

$$D_{\mathrm{Rao}}[p_{\theta_1}, p_{\theta_2}] = \rho_g(\theta_1, \theta_2) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\,\mathrm{d}t, \gamma(0) = \theta_1, \gamma(1) = \theta_2$$

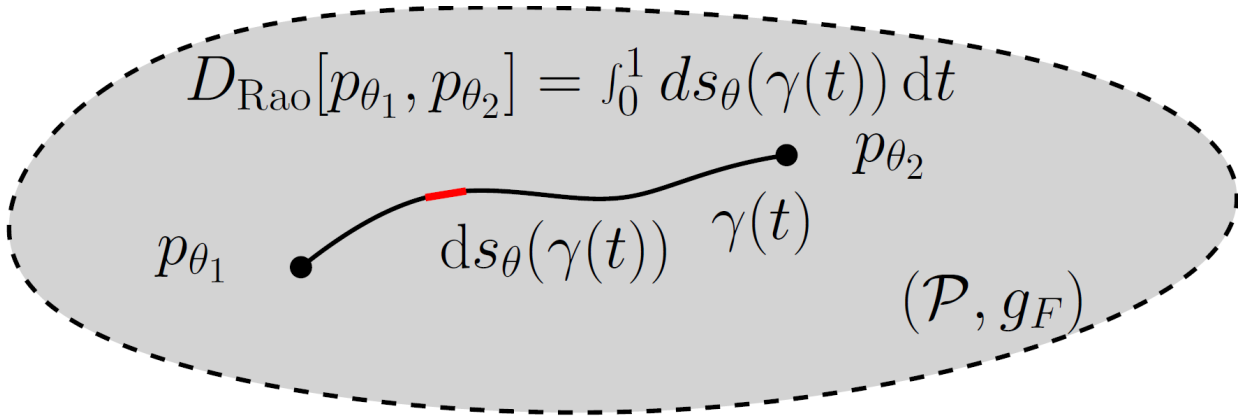
$$= \int_0^1 ds_\theta(\gamma(t))\,\mathrm{d}t$$

où γ est la géodésique Riemannienne
(sinon rajouter un min sur tous les chemins γ)

élément de longueur infinitesimal

$$\mathrm{d}s_\theta^2(t) = \sum_{i=1}^{D}\sum_{j=1}^{D} g_{ij}(\theta)\dot{\theta}_i(t)\dot{\theta}_j(t)$$

$$\dot{\theta}_k(t) = \frac{d}{\mathrm{d}t}\theta_k(t)$$



En pratique :

- on doit calculer les géodésiques qui sont caracterisées localement comme les **courbes de longueurs minimisantes** en géométrie Riemannienne.
- C'est une tâche pour laquelle on ne connait toujours pas de solution pour un modele statistique usuel comme les **lois normales multivariées** !

Invariance de la distance de Rao par reparamétrage

Considérons deux paramétrages du modèle statistique :

$$\mathcal{P} = \{p_\theta \ : \ \theta \in \Theta\} = \{p_\eta \ : \ \eta \in H\}$$

La matrice d'information de Fisher est **covariante** par reparamétrage :

$$I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial\theta_i}{\partial\eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial\theta_i}{\partial\eta_j}\right]$$

... mais l'élément infinitesimal est lui **invariant** : $\mathrm{d}s_\theta = \mathrm{d}s_\eta$

... si bien que la distance de Rao est **invariante** : $\rho_{\mathrm{Rao}}(p_{\eta_1}, p_{\eta_2}) = \rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})$

- C'est le **premier principe d'invariance en géométrie de l'information**

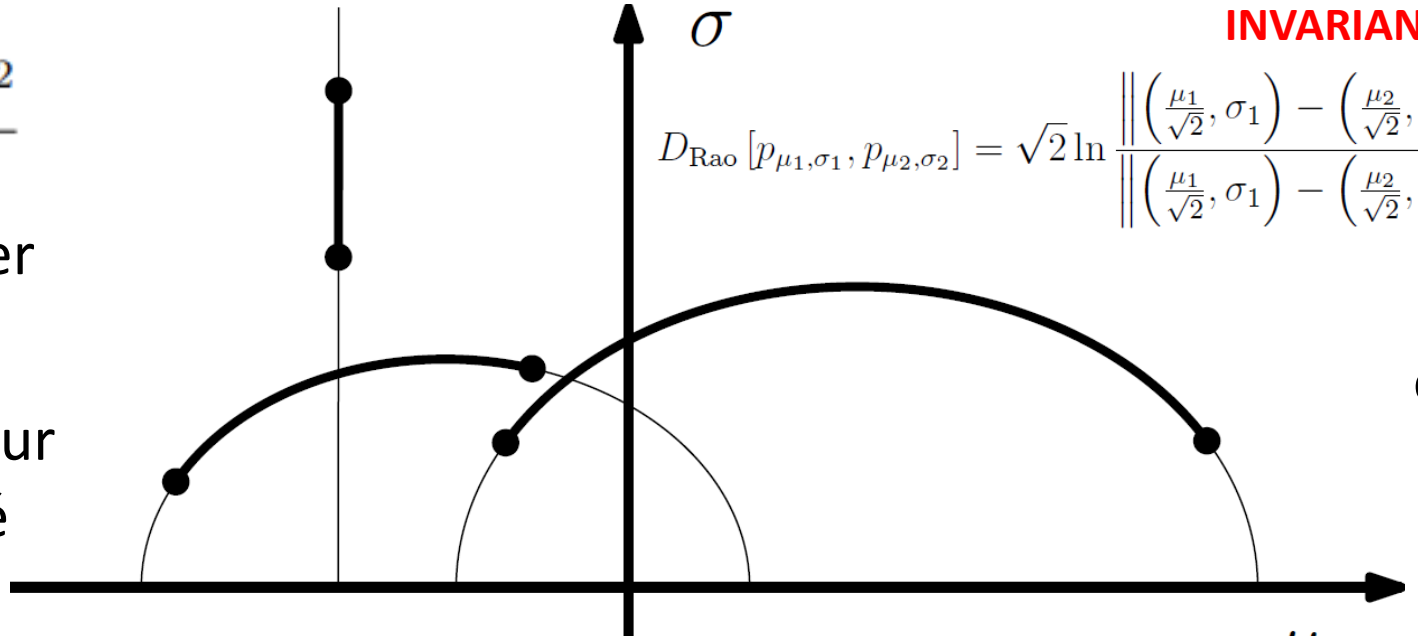
Géométrie de Fisher-Rao des lois normales

INVARIANT

$$ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$

métrique de Fisher

Demi-plan supérieur de Poincaré étiré

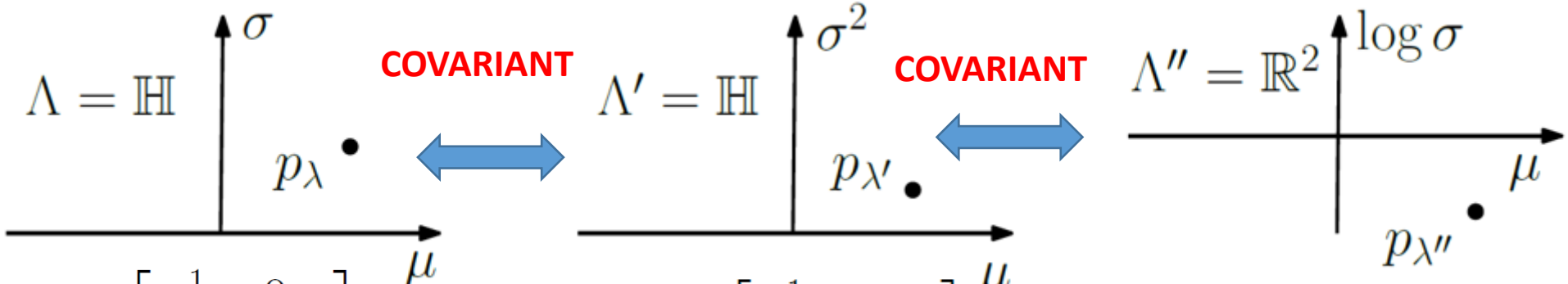


INVARIANT

$$D_{\mathrm{Rao}}\left[p_{\mu_1,\sigma_1}, p_{\mu_2,\sigma_2}\right] = \sqrt{2}\ln \frac{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| + \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| - \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}$$

distance de Rao ou de Fisher-Rao

Différents domaines de paramétrage & MIF



$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

$$I(\mu,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

$$I(\mu,\log\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2 \end{bmatrix}$$

En général, les familles de lois positions-échelles ont une géométrie de Fisher hyperbolique

Variété de Fisher-Rao intrinsèque et extrinsèque

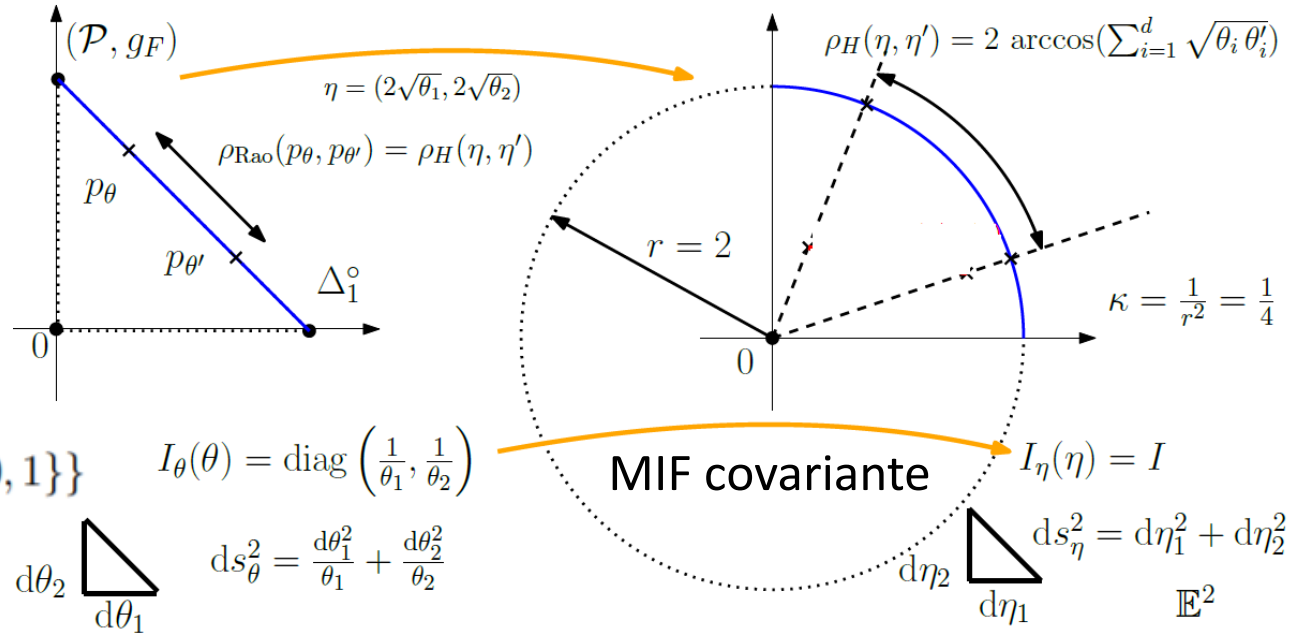
- Une variété Riemannienne de dimension D peut être visualisée comme une **surface** d'un espace Euclidien de dimension $O(D^2)$:

  Plongement isométrique de la variété

- Par exemple, la distance de Rao entre deux lois de Bernoulli ou catégorielles se trouve facilement en plongeant le simplexe standard sur la sphère de rayon 2 par la transformation 2√.

Espace des paramètres
Variété de Fisher
intrinsèque
de dimension 1
(famille de Bernoulli)

$$\mathcal{P} = \{p^x(1-p)^{1-x}, p \in (0,1), x \in \{0,1\}\}$$



Variété de Fisher
extrinsèque
plongée dans
R^2

Les neurovariétés et l'apprentissage profond

- Un réseau de neurones artificiels (ex., perceptrons multi-couches) est décrit par une architecture utilisant un très grand nombre de paramètres θ
- On considère des **réseaux de neurones (NN) stochastiques** avec une réponse bruitée :

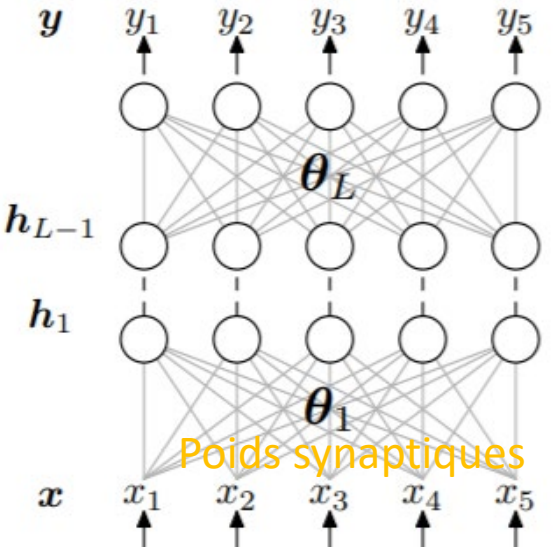
$$y = \mathrm{NN}_\theta(x) + \epsilon$$

... avec un bruit Gaussien :

$$p_\theta(x,y) = p(x)p_\theta(y|x) = \frac{p(x)}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y-\mathrm{NN}_\theta(x))^2\right)$$

- La neurovariété est

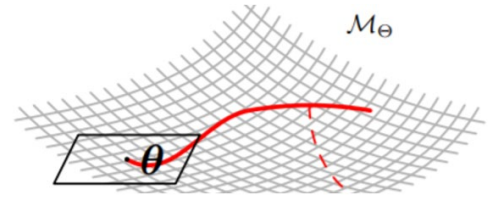
$$\mathcal{P} = \{p_\theta(x,y) \ : \ \theta \in \Theta\}$$



réseau de neurones artificiels

neurovariété

- Maximiser la vraisemblance revient à **minimiser la moyenne des erreurs quadratiques**
- Étant donné un jeu de données d'entrainement, on doit apprendre les paramètres du réseau par une methode de **descente de gradient**. On visualise l'apprentissage dynamique par une **trajectoire** sur cette neurovariété. On observe des **phénomènes de plateaux** lorsqu'on approche une **singularité** où la matrice de Fisher n'est pas de rang plein.



trajectoire d'apprentissage

[SN 2017]

Le gradient naturel : Plus forte pente Riemannienne

Descente de gradient ordinaire pour minimiser une **fonction de perte** E(.) :

$$\theta_{t+1} = \theta_t - \alpha \nabla E(\theta_t)$$

Pas d'apprentissage α

- dépend du choix de paramétrage
- phénomènes de plateaux près des singularités (MIF proche d'être dégénerée)

Le gradient naturel est invariant au reparamétrage et ne fait pas de plateaux :

$$\tilde{\nabla} E(\theta) := G(\theta)^{-1} \nabla_\theta E(\theta)$$

$$\tilde{\nabla} E_\eta(\eta) = \tilde{\nabla} E_\theta(\theta)$$

Descente par gradient naturel :

$$\theta_{t+1} = \theta_t - \alpha \tilde{\nabla} E(\theta_t)$$

où α est le pas d'apprentissage

C'est différent d'une méthode de descente par gradient Riemannien qui se base sur l'exponentielle Riemannienne qui est coûteuse en temps de calcul.

Qu’est-ce que la géométrie de l’information ? (4/4)

- Une **structure duale Riemannienne** permet d'expliquer la dualité entre l'**estimateur** par maximum de vraisemblance et la **famille de modèles statistiques** issus du principe de l'entropie maximale : Explique le lien entre l'entropie de Shannon, la divergence de Kullback-Leibler et les familles exponentielles en statistique.

- **Second principe d'invariance** par **statistique suffisante** $p_\lambda(x)$

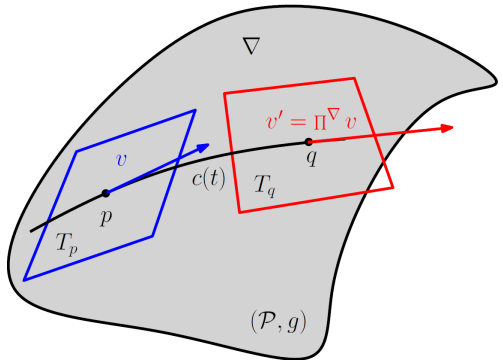
$x \in$ univers Ω

Cette structure duale :

- Ouvre de **nouvelles perspectives** : Par exemple, entropies non-extensives, systèmes complexes et géométries conformales (des familles exponentielles déformées), etc.
- A de nombreux domaines d'applications de la GI : traitement du signal (Radar, interfaces cerveau-machine, etc.), imagerie médicale, apprentissage, IA, etc.

Les géodésiques dépendent de connexions affines

- On a vu que les géodésiques sont localements des plus courts chemins en géométrie Riemannienne
- Géodésique γ(t) définie par ∇ est une courbe **∇-autoparallèle** :



$$\nabla_{\dot\gamma}\dot\gamma = 0, \qquad \dot\gamma = \frac{d}{\mathrm{d}t}\gamma(t)$$

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p}\Gamma_{ij}^{k}\frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,$$

avec $\nabla_X T$ l'operateur de **différentiation covariant** pour un champ de vecteurs X

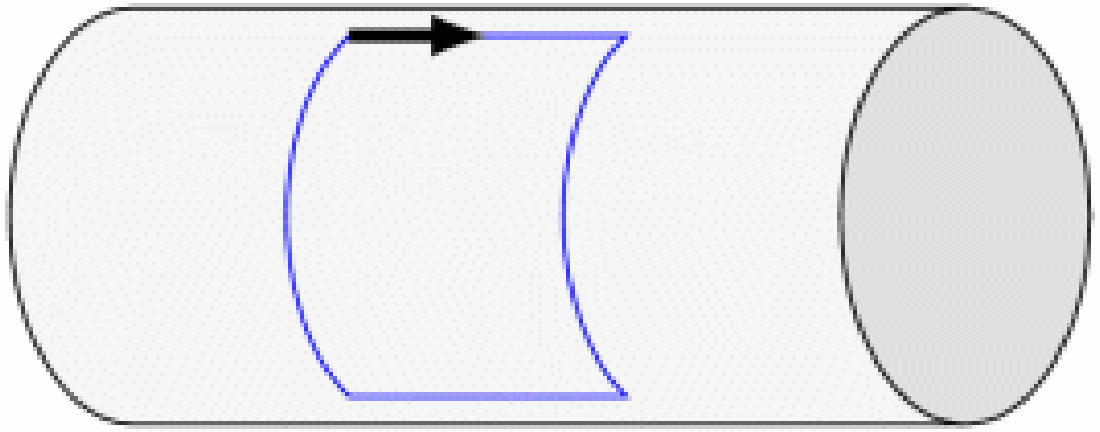
Les D^3 **symboles de Christoffel** Γ sont des fonctions qui caractérisent la connexion affine ∇

- En géométrie Riemanienne, la connexion par défaut est celle de **Levi-Civita** qui est construite à partir du tenseur métrique g (donc implicite en géo. Rie.) :

$$\nabla = {}^{g}\nabla$$

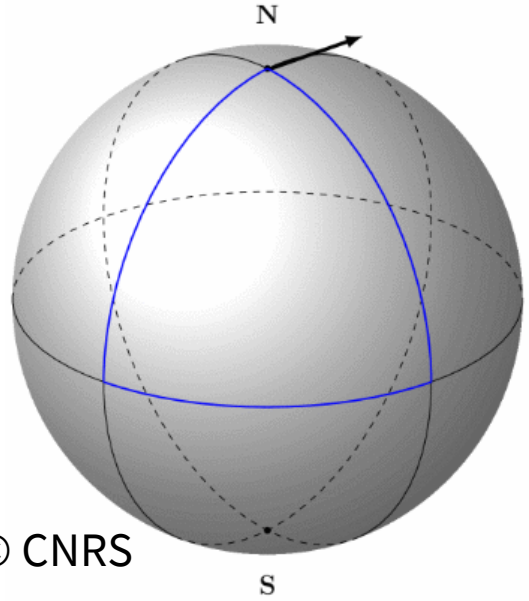
$$\Gamma_{ij}^{k} = \frac{1}{2}\sum_{m=1}^{p}\left(\frac{\partial g_{im}(\boldsymbol{\theta})}{\partial\theta_j} + \frac{\partial g_{jm}(\boldsymbol{\theta})}{\partial\theta_i} - \frac{\partial g_{ij}(\boldsymbol{\theta})}{\partial\theta_m}\right) g^{mk}(\boldsymbol{\theta}), \quad i, j, k = 1, \ldots, p,$$

Connexion affine ∇ : Visualisation de la courbure par le transport parallèle sur des boucles infinitesimales



© CNRS

Le cylindre est plat : courbure 0



© CNRS

© wikipedia

Élie Cartan
1869-1951

La sphère est de courbure constante >0

Une connection est plate s'il existe un systeme local de coordonnées θ pour lequel les symboles de Christoffel sont tous nuls : **Γ(θ)=0**

→ Les géodésiques sont tracées comme des **segments de droite** dans la carte θ

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,$$

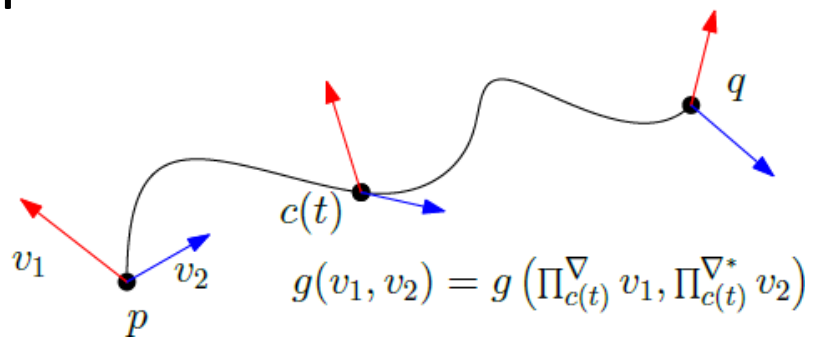
Segment de droite

La nature duale de la géométrie de l'information

(M, g, ∇, ∇^*) tel que $\frac{\nabla + \nabla^*}{2} = {}^g\nabla$

- Étant donné une connexion affine sans torsion ∇ et une métrique $\mathbf{g}$, on peut construire une **unique connexion duale ∇^* sans torsion** telle que la métrique est préservée par le bi-transport parallèle :

$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)}.$$



- Le dual de la connexion duale est la connexion primale : $(\nabla^*)^* = \nabla.$
- Comment trouver des connexions duales ?
  - Méthode d'Amari-Nagaoka (1982) : les **α-connexions** (Chentsov 1972)
  - Méthode d'Eguchi (1983) : construit les connexions duales à partir de divergences

L' α-géométrie duale d'Amari et Nagaoka

Structure $(\mathcal{P}, g_F, \nabla^\alpha, \nabla^{-\alpha})$

∇^α définit par les symboles de Christoffel $\Gamma^\alpha_{ij,k} = E_\theta\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right]$

Quelques α-connexions:

- **0-connexion** = **connexion métrique de Fisher Levi-Civita : Variété Fisher-Rao**
- **1-connection** est appellée la **connexion exponentielle** **[Efron 1975]**
- **-1 connection** est appellée la **connexion de mélange** **[Dawid 1975]**

$$(\nabla^e)^* = \nabla^m \qquad (\nabla^m)^* = \nabla^e$$

$$\nabla^\alpha = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m$$

La **géométrie duale em** va être utilisée pour étudier la dualité estimateurs/modèles

La géométrie d'Eguchi induite par une divergence

- Structure $(M, {}^{D}g, {}^{D}\nabla, {}^{D}\nabla^{*})$
- On obtient une **divergence** (fonction de contraste) à partir d'une distance en statistique entre lois parametriques d'une famille. Par exemple, pour l'entropie relative entre deux lois d'une famille P, on a

divergence paramétrique
Divergence de contraste

$$D_{\mathrm{KL}}^{\mathcal{P}}(\theta_1 : \theta_2) := D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}]$$

divergence statistique

- La métrique d'Eguchi associée à D est ${}^{D}g_{ij}(\theta_1) = \partial_{\theta_1}\partial_{\theta_2} D(\theta_1 : \theta_2)|_{\theta_1=\theta_2}$
- La connexion d'Eguchi associée à D est ${}^{D}\Gamma_{ij,k} = g\left({}^{D}\nabla_{\partial_i}\partial_j, \partial_k\right) = -\partial_{\theta_1^i}\partial_{\theta_1^j}\partial_{\theta_2^k} D\left(\theta_1 \| \theta_2\right)\Big|_{\theta_1=\theta_2}$
- On définit la **divergence duale** en changeant l'ordre des paramètres :

$$D^{*}(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$$

- On obtient les connexions duales ${}^{D}\nabla^{*} = {}^{D^{*}}\nabla$ et on a

Connexion de Levi-Civita

$${}^{D}g\nabla = \frac{1}{2}\left({}^{D}\nabla + {}^{D^{*}}\nabla\right)$$

Les f-divergences et leurs connexions associées

- L'entropie relative ou divergence de Kullback-Leibler n'est qu'un exemple d'une famille de divergences : les **f-divergences** **[Csiszar'63] [Ali&Silvey'66]**

$$I_f[p:q] = \int p f(q/p) \mathrm{d}\mu \Rightarrow D_{\mathrm{KL}}[p:q] = \int p \log p/q \mathrm{d}\mu = I_{f_{\mathrm{KL}}}[p:q] \quad f_{\mathrm{KL}}(u) = -\log u$$

Distance séparable

- Le générateur f(.) est convexe, strictement convexe à 1. On peut fixer f'(1)=0 et f''(1)=1 pour obtenir une **f-divergence standard**.
- La **f-divergence duale** est $I_f^*[p:q] = I_f[q:p] = I_{f^*}[p:q]$ avec $f^*(u) = uf(1/u)$
- Métrique induite par la méthode d'Eguchi est celle de Fisher : ${}^{I_f^{\mathcal{P}}}g = g_F = {}^{I_{f^*}^{\mathcal{P}}}g$
- Les **f-connexions** induites par Eguchi pour les f-divergences entre lois d'une famille P coïncident avec les **α-connexions** d'Amari et de Nagaoka :

$$I_f^{\mathcal{P}}\nabla = {}^{\mathcal{P}}\nabla^{\alpha_f}$$

$$\alpha_f = 3 + 2\frac{f'''(1)}{f''(1)}$$

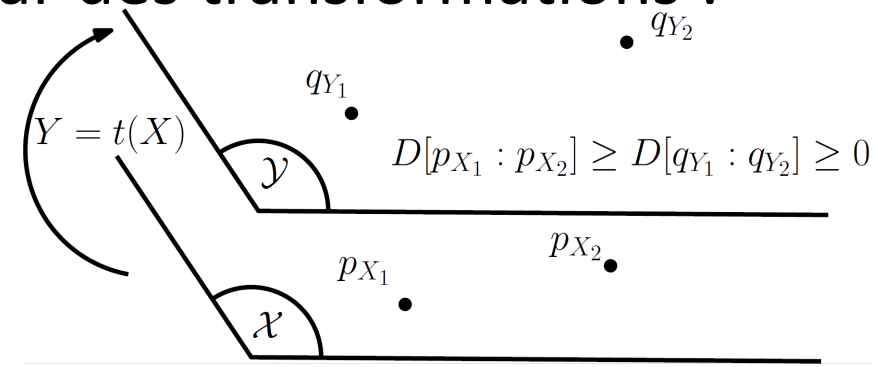
Les distances en statistique et leurs monotonicités

- On considère une transformation Y=t(X) de variables aléatoires entre deux espaces mesurables :

$$t: \ (\mathcal{X}, \Sigma) \to (\mathcal{Y}, \Sigma') \qquad Y_i = t(X_i)$$

- **Deuxième principe d'invariance** : On ne doit pas pouvoir augmenter la discrimination d'une divergence en statistique par des transformations :

$$D[p_{X_1} : p_{X_2}] \geq D[q_{Y_1} : q_{Y_2}] \geq 0$$



- monotonicité de la MIF : $I_{t(X)}(\theta) \leq I_X(\theta)$
- On a égalité que si et seulement si t(X) est une **statistique suffisante**.
- Une statistique suffisante résume exhaustivement toute l'information nécessaire pour faire l'inférence du paramètre θ : $\Pr(x|\theta) = \Pr(x|t)$
- **Les f-divergences sont les seules distances monotones séparables**

Les familles exponentielles ont une statistique suffisante

- Une famille exponentielle a ses densités qui s'expriment canoniquement comme : $p_\theta(x) = \exp\left(\langle \theta, t(x) \rangle - F(\theta)\right)$ μ (eg., mesure de Lebesgue ou de comptage)

Ici, c'est le produit scalaire Euclidien

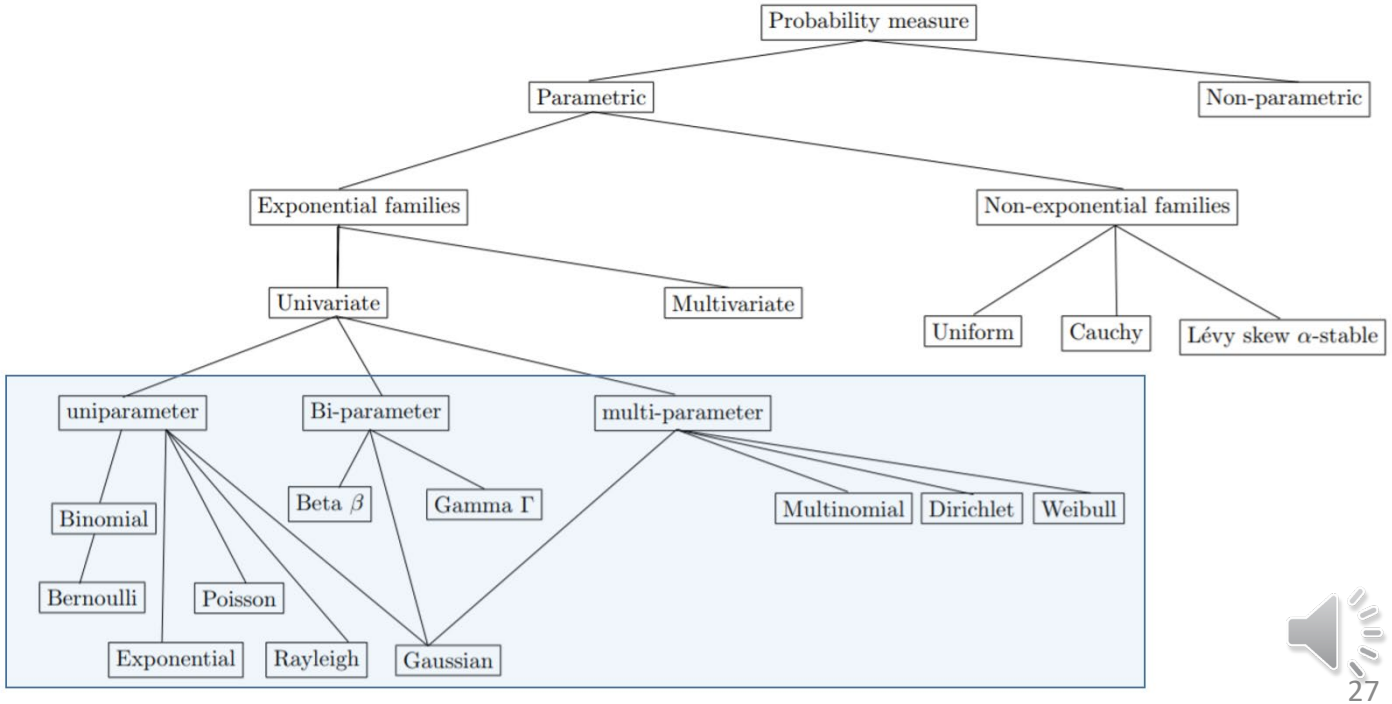
où F is est une **fonction analytique strictement convexe et differentiable**:

$$F(\theta) = \log \int \exp(\theta x) \mathrm{d}\mu(x)$$

F : fonction log partition ou fonction cumulante

Espace des paramètres naturels :

$$\Theta = \left\{\theta \ : \ \int \exp(\theta x) \mathrm{d}\mu(x) < \infty\right\}$$

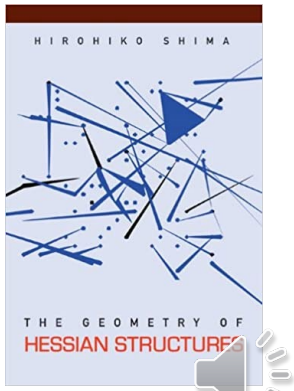


Géométries duallement plates des familles exponentielles

- Modèle statistique : famille exponentielle $\mathcal{P} = \{p_\theta(x) = \exp(x^\top \theta - F(\theta))\}$
  ou plus généralement $\mathcal{P} = \{\exp(\theta^\top t(x) - F(\theta))\, h(x)\}$ $p_\theta(x) = p^\eta(x)$
- La **connexion exponentielle** et la **connexion duale de mélange** sont **plates** : **Espaces duallement plats des familles exponentielles** : $(M, g_F, \nabla^e, \nabla^m)$

- La matrice de Fisher est hessienne $g_F(\theta) = I(\theta) = \mathrm{Cov}[t(X)] = \nabla^2 F(\theta)$
- Par la transformation de **Legendre-Fenchel,** on obtient un système de coordonnées dual $F^*(\eta) = \sup_\theta \theta^\top \eta - F(\theta), \eta = \nabla F(\theta) = E[t(X)]$
- Paramétrage par les moments : $\boxed{p_\theta(x) = p^\eta(x)}$
- La matrice de Fisher exprimée avec le parametre moment:

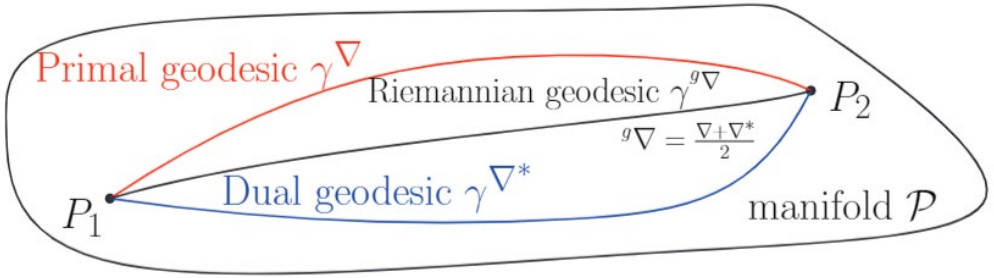
$$I(\eta) = \nabla^2 F^*(\eta) = g_F(\theta)^{-1}$$

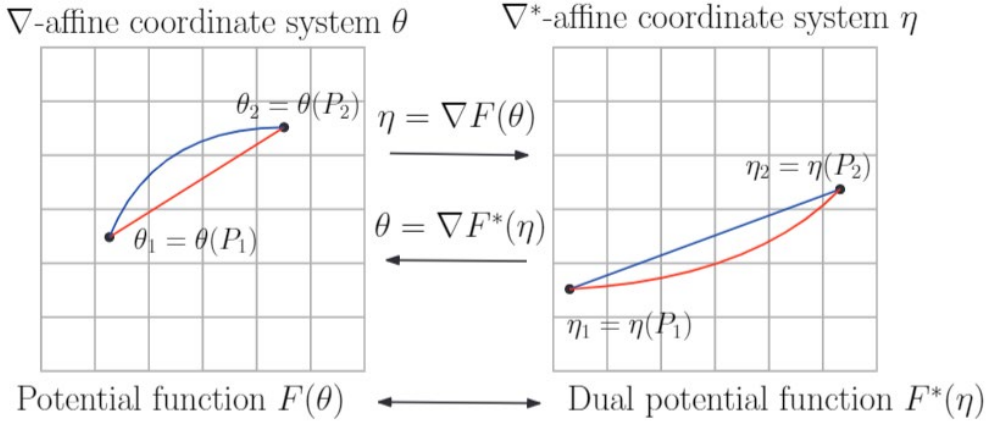


Géométries duallement plates: Structures Hessiennes

- Les géodésiques primales/duales sont des **segments de droite** dans le système des paramètres naturels/moments :

$$p_\theta(x) = p^\eta(x)$$





géodésique primale $\gamma_{p_{\theta_1} p_{\theta_2}}(t) = p_{(1-t)\theta_1 + t\theta_2}$

géodésique duale $\gamma^*_{p_{\theta_1} p_{\theta_2}}(t) = p^{(1-t)\eta_1 + t\eta_2}$

- Dans les espaces duallement plats, on a une **divergence canonique** : la **divergence de Bregman** qui revient à calculer la **divergence duale de Kullback-Leibler** entre les densités correspondantes :

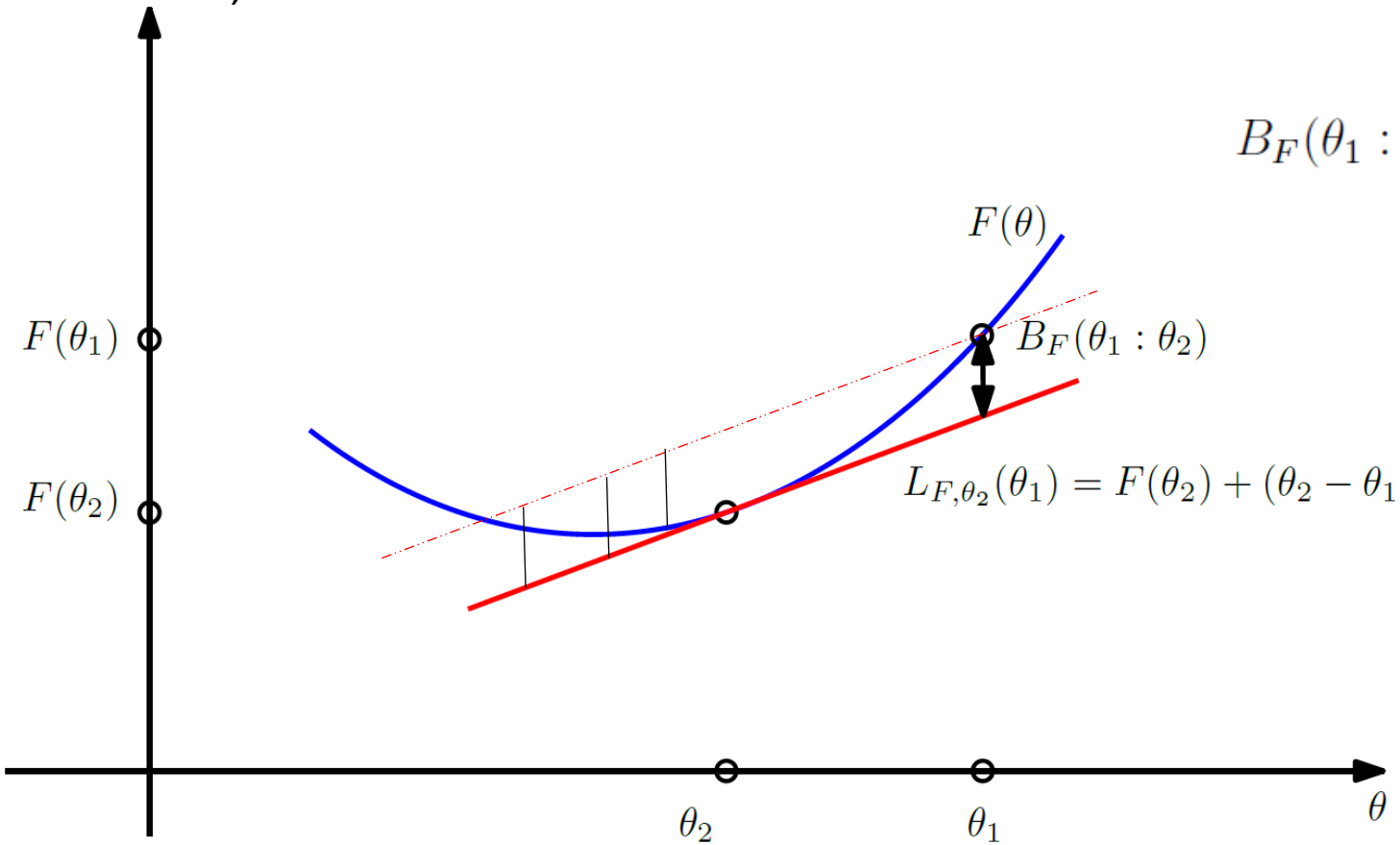
$$B_F(\theta_1 : \theta_2) = D_{\mathrm{KL}}{}^*[p_{\theta_1} : p_{\theta_2}] = D_{\mathrm{KL}}[p_{\theta_2} : p_{\theta_1}]$$

La divergence de Bregman illustrée

- Soit F(θ) une fonction strictement convexe et différentiable définie dans un domaine ouvert Θ
- La divergence de Bregman se lit comme la distance verticale entre le point $(\theta_1, F(\theta_1))$ et l'approximation linéaire de F(θ) à θ_2 évaluée à θ_1 :

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - \underbrace{\left(F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2)\right)}_{L_{F,\theta_2}(\theta_1)}$$

$$= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)$$



[Bregman 1967]

Les Coordonnées mixtes et la divergence de Fenchel-Young

- Paramétrage dual $\theta = \nabla F^*(\eta) \iff \eta = \nabla F(\theta)$
- La fonction conjuguée de F est $F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta))$
- **Inégalité de Fenchel-Young** :

$$F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2$$

avec égalité si et seulement si $\eta_2 = \nabla F(\theta_1)$

$\nabla F^* = (\nabla F)^{-1}$

gradients réciproques des fonctions conjuguées

- La **divergence de Fenchel-Young** utilise les coordonnées mixtes θ et η pour exprimer la divergence de Bregman $B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2)$:

$$Y_{F,F^*}(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 = Y_{F^*,F}(\eta_2, \theta_1)$$

Les divergences duales de Bregman et de Fenchel-Young

- **Identité des divergences de Bregman duales** : $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$

  (la divergence de Bregman duale est la divergence renversée pour le générateur dual)

- La divergence duale est la divergence renversée : $D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$

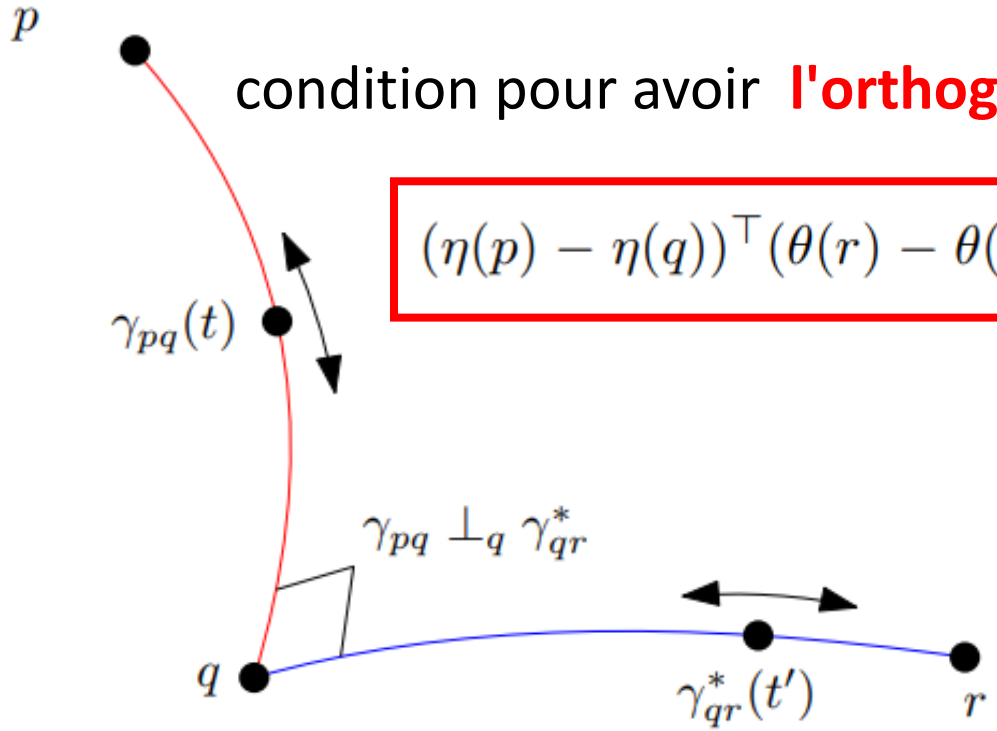
- Paramétrages primal, dual, ou mixte pour la divergence canonique des espaces duallement plats :

$$B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2) = Y_{F^*,F}(\eta_2, \theta_1) = B_{F^*}(\eta_2 : \eta_1)$$

Sur une variété de Bregman, on peut donc avoir 2^n formules equivalents a n termes

Théorème de Pythagore des espaces duallement plats

Interprétation avec un théorème de Pythagore



condition pour avoir **l'orthogonalité** :

$$(\eta(p) - \eta(q))^\top (\theta(r) - \theta(q)) = 0$$

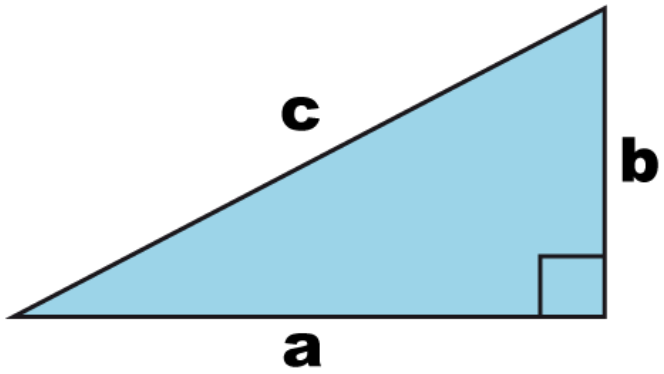
$$D_F(\gamma_{pq}(t) : \gamma_{qr}(t')) = D_F(\gamma_{pq}(t) : q) + D_F(q : \gamma^*_{qr}(t')), \quad \forall t, t' \in (0,1).$$

Théorème de Pythagore

géométrie Euclidienne auto-duale

$$F_{\mathrm{Eucl}}(\theta) = \frac{1}{2}\theta^\top\theta \qquad g_{F_{\mathrm{Euc}}} = I$$

$$B_{F_{\mathrm{Eucl}}}(\theta_1 : \theta_2) = \frac{1}{2}\rho^2_{\mathrm{Eucl}}(\theta_1, \theta_2)$$



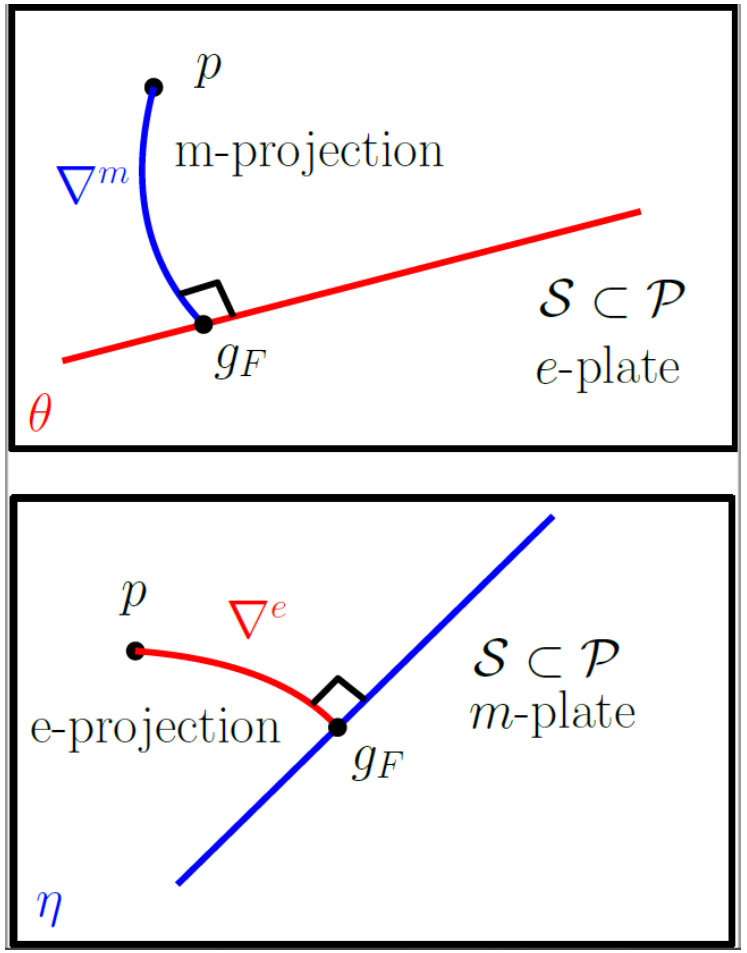
a² + b² = c²

Propriété de la divergence de Bregman à trois paramètres :

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3)) \geq 0$$

Théorèmes sur les projections d'information

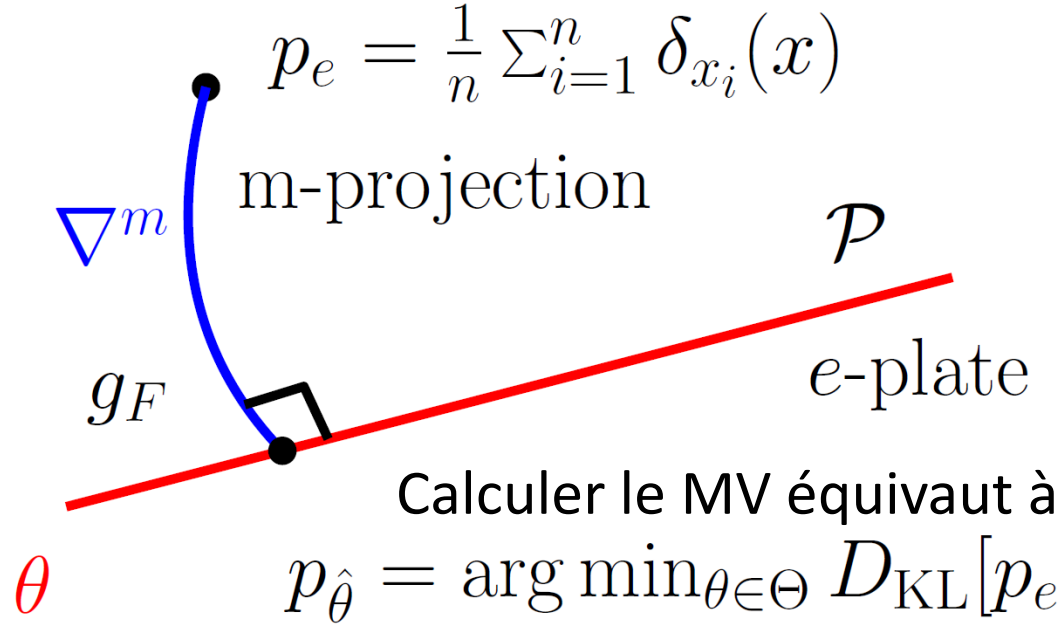
- On définit la **e-projection** et la **m-projection** d'un point sur une sous-variété en fonction des connexions affines ∇^e (∇^{+1}) et ∇^m (∇^{-1}) et de l'orthogonalité donnée par la métrique de Fisher g_F
- Une sous-variété est **e-plate** si et seulement si elle est représentée dans le système de coordonnées θ par un sous-espace affine. Idem pour une sous-variété **m-plate** vis-à-vis de η
- Le théorème de Pythagore permet de démontrer que la e-projection d'un point est **unique** sur une sous variété m-plate et correspond à la minimisation d'une divergence de Bregman. Idem pour la m-projection d'un point sur une sous-variété e-plate qui est obtenue par la minimisation de la divergence duale de Bregman.



Maximum de vraisemblance et m-projection

Un échantillon $\{x_1, ..., x_n\}$ i.i.d. d'une loi appartenant à une **famille exponentielle** P (**e-plate**)

La distribution empirique est appellée le **point observé**



Calculer le MV équivaut à

$$p_{\hat{\theta}} = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}[p_e : p_\theta]$$

Maximum de vraisemblance :

$$p_{\hat{\theta}} = \mathrm{Proj}_{\mathcal{P}}^{\nabla^m}(p_e)$$

Divergence canonique :

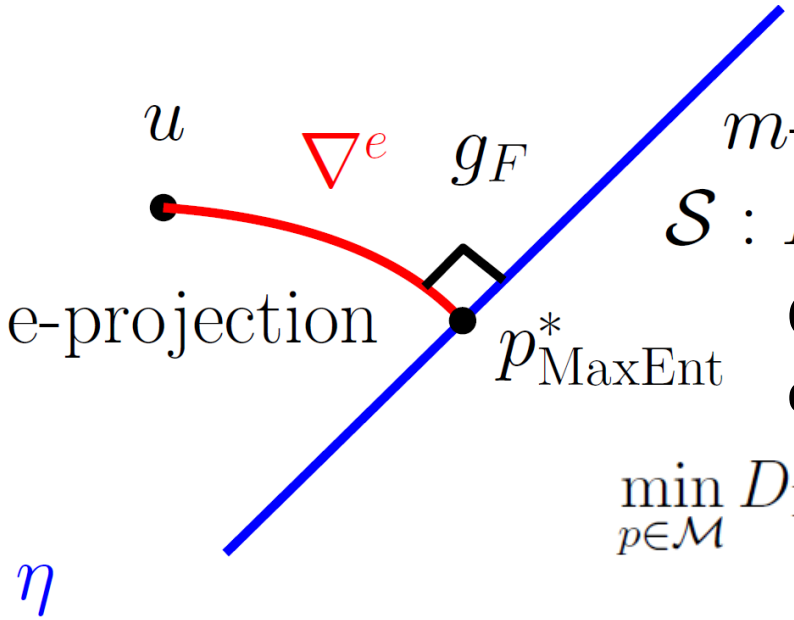
$$D_{\nabla^m} = D_{\mathrm{KL}}$$

En remplaçant la divergence de Kullback-Leibler par une divergence arbitraire D[.:.], on définit ainsi des **D-estimateurs**. MV est le KLD-estimateur.

Maximum d'entropie et e-projection

- Etant donné des **observations** $E[t_i(x)]=m_i$, le **principe du maximum d'entropie** de Jaynes estime *la distribution* qui maximise l'entropie de Shannon et qui satisfait les contraintes de moment.

$$E_p[t_1(X)] = m_1, \ldots, E_p[t_D(X)] = m_D$$
$$t(x) = (t_1(x), \ldots, t_D(x))$$
$$m = (m_1, \ldots, m_D)$$



Calculer la distrib. MaxEnt équivaut à

$$\min_{p \in \mathcal{M}} D_{\mathrm{KL}}[p : u] = \min_{p} D^*_{\mathrm{KL}}[u : p]$$

$$p^*_{\mathrm{MaxEnt}} = \mathrm{Proj}^{\nabla^e}_{\mathcal{S}}(u)$$

Divergence canonique duale :

$$D_{\nabla^e} = D^*_{\mathrm{KL}}$$

- Les distributions maximisant l'entropie les contraintes $E[t(x)]=\eta$ pour tous les η forment une **famille exponentielle**: $p^* \in \mathcal{E} := \{p_\theta(x) = \exp(\sum t_i(x)\theta_i - F(\theta))\}$
- Par exemple, les distributions MaxEnt pour $E[x]=\eta_1$ et $E[x^2]= \eta_2$ forment la famille des lois normales (univariée d'ordre 2)

Les projections alternées : l'algorithme em (= $\nabla^e \nabla^m$)

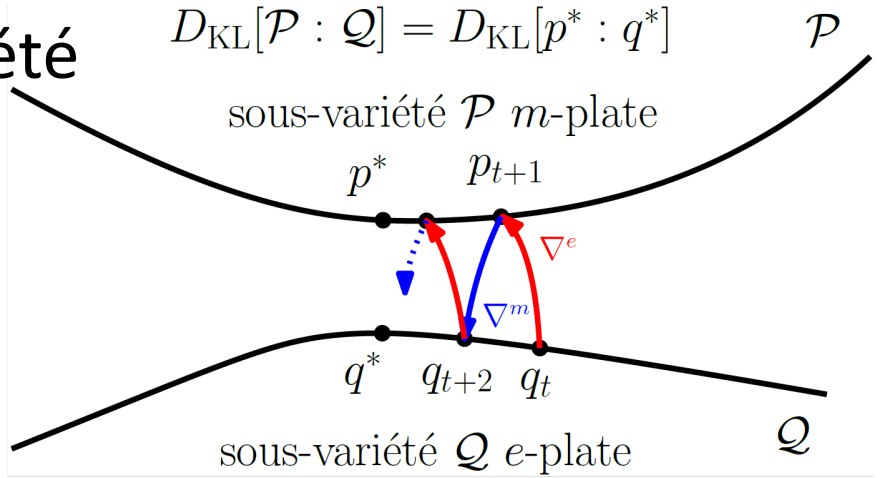
- Trouver la **distance minimale entre deux sous-variétés**
  = Résoudre la minimisation conjointe

$$\min_{p\in\mathcal{P}} \min_{q\in\mathcal{Q}} D_{\mathrm{KL}}[p:q]$$

- Lorsque la sous-variété P est **m-plate** et la sous-variété Q est **e-plate**, alors on a unicité des e/m-projections alternées suivantes en partant initialement de q_1 :

e-projection : $p_{t+1} = \arg\min_{p\in\mathcal{P}} D_{\mathrm{KL}}[p:q_t]$

m-projection : $q_{t+2} = \arg\min_{q\in\mathcal{Q}} D_{\mathrm{KL}}[p_{t+1}:q]$



- On converge vers les deux points qui minimisent la divergence de Kullback-Leibler entre P et Q :

$$D_{\mathrm{KL}}[p^*:q^*] = \min_{p\in\mathcal{P}} \min_{q\in\mathcal{Q}} D_{\mathrm{KL}}[p:q] = \lim_{t\to\infty} D_{\mathrm{KL}}[p_{t+1}:q_t]$$

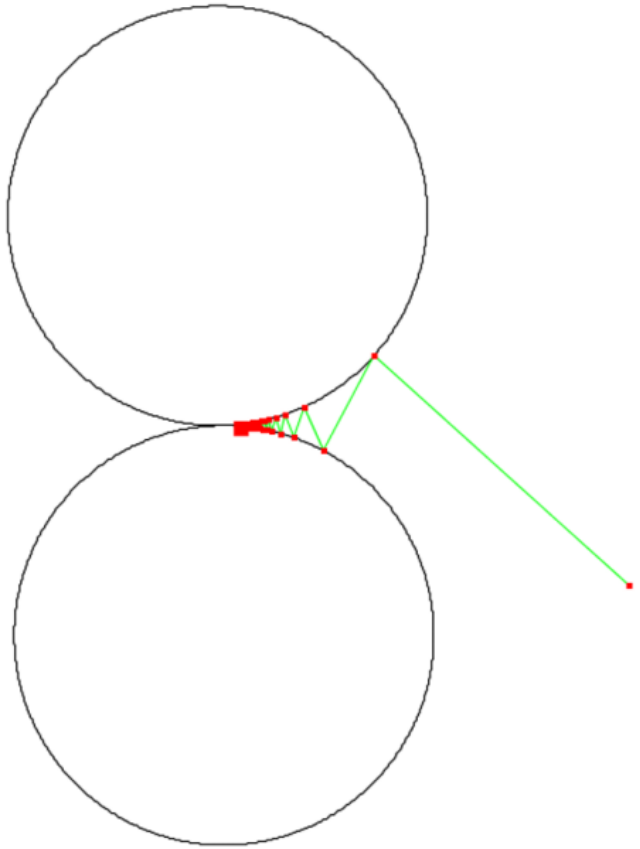
L'algorithme em est utile pour :
- l' interprétation de EM en statistique
- l'analyse des modèles génératifs profonds comme les VAEs ou GANs

Les projections cycliques de Bregman (dans une carte)

- Soit n **objets convexes** O_1, ..., O_n dans un système de coordonneés θ sur un domaine convexe Θ
- On désire trouver un point dans l'intersection commune de ces objets si elle existe
- On répète <u>cycliquement</u> les **projections de Bregman**

$$\theta_0 \in \Theta, t \leftarrow 0$$

$$\theta_{t+1} = \arg \min_{\theta \in O_{1+(t \mod n)}} B_F(\theta_t : \theta)$$



- La séquence converge **vers un point de l'intersection si elle existe**.

L'information de Chernoff et les Tests d'hypothèses

- Soit deux distributions P_1 et P_2 , et n échantillons i. i. d. x_1, ..., x_n qui proviennent du modèle de mélange 1/2 P_1 + 1/2 P_2
- **Quelle règle pour classer ces n échantillons en étiquettes P1 ou P2 ?**
- La meilleure règle est celle du **maximum a posteriori** (MAP)
- La **probabilité d'erreur** est bornée par $P_e^n = 2^{-nC(P_1,P_2)}$

  avec C qui désigne est la **divergence de Chernoff**

$$C(P,Q) = -\log \min_{\alpha\in(0,1)} \int p^{\alpha}(x) q^{1-\alpha}(x) \mathrm{d}\nu(x).$$

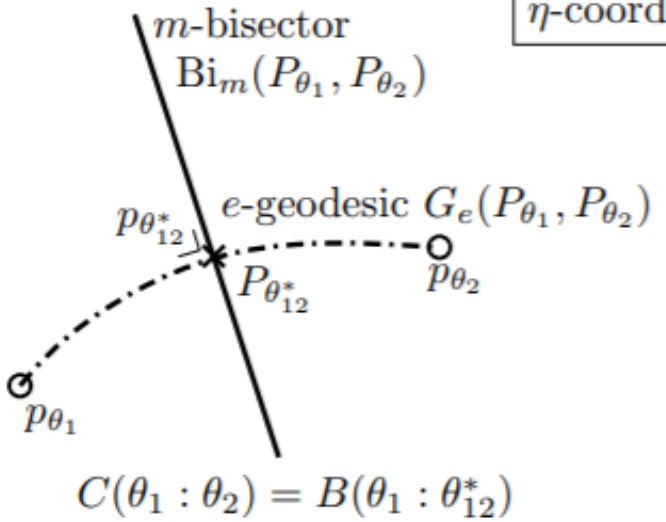
Quand P_1 et P_2 sont deux distributions d'une meme famille exponentielle :

$$C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})$$

α* est l'exposant optimal dans (0,1)

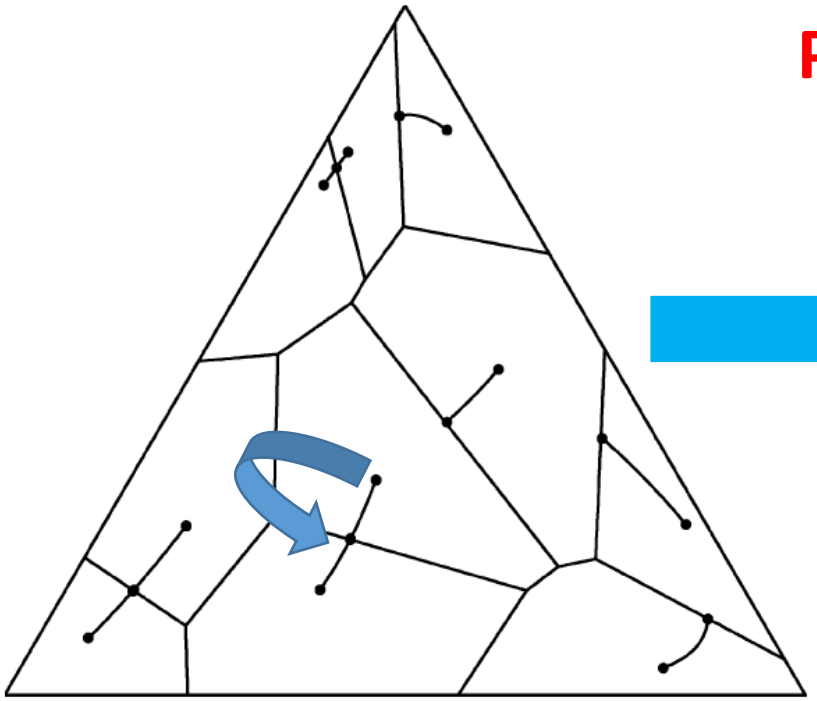
$$P^* = P_{\theta_{12}^*} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$$

η-coordinate system



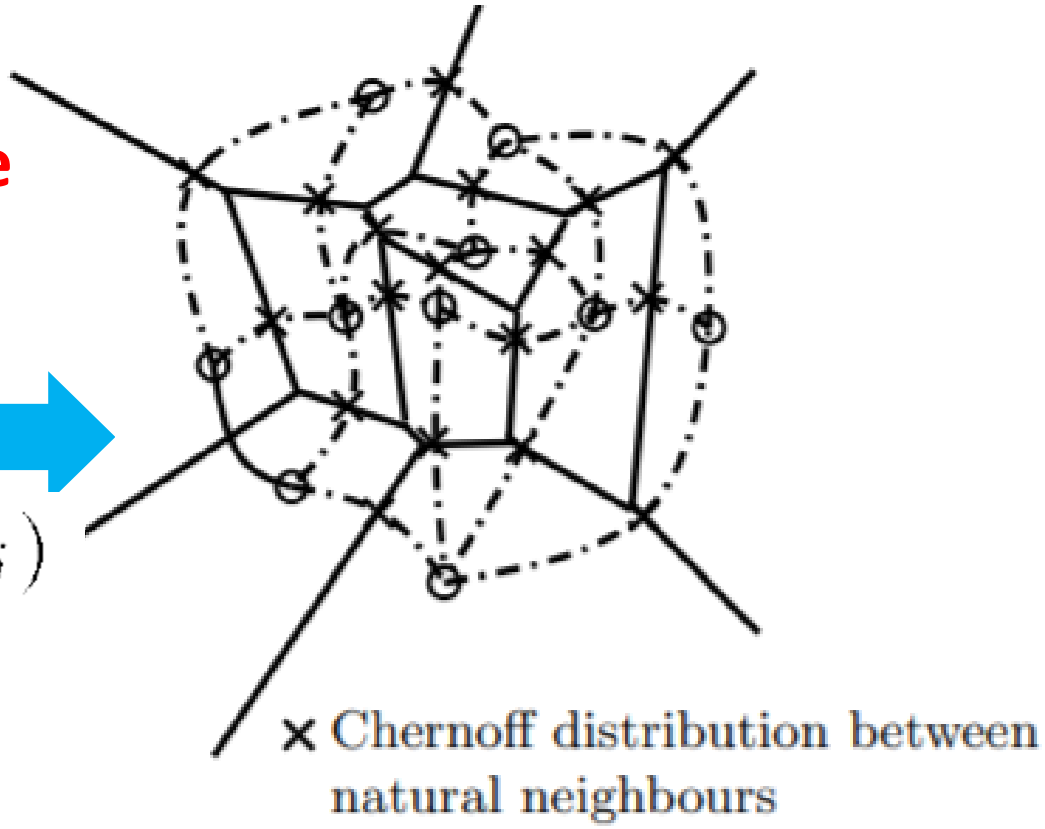
L'information de Chernoff pour plusieurs hypothèses

Probabilité d'erreur : $P_e^n = 2^{-nC(P_i^*, P_j^*)}$



Paire la plus proche pour l'information de Chernoff

$$\underset{i \neq j}{\operatorname{argmin}}\, C(P_i, P_j)$$



Simplexe standard (lois categorielles)

diagramme de Voronoï pour la divergence de Kullback-Leibler

Variété d'une famille exponentielle

diagramme de Voronoï Bregmannien

Géométrie algorithmique : diagramme de Voronoï Bregmannien

Gradient naturel dans les espaces duallement plats

Sur une variété Hessienne globale (variété de Bregman induite par une fonction convexe F), la matrice d'information de Fisher peut s'exprimer comme

$$I_\theta(\theta) = \nabla_\theta^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$$

Définition du paramètre moment

Définition du **gradient naturel** par rapport à θ :

$$\tilde{\nabla}_\theta L_\theta(\theta) := I_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta)$$

$$= (\nabla_\theta \eta)^{-1} \nabla_\theta \eta \nabla_\eta L_\eta(\eta)$$

Définition du **gradient ordinaire** par rapport à η

$$= \nabla_\eta L_\eta(\eta)$$

dérivation des fonctions composées :

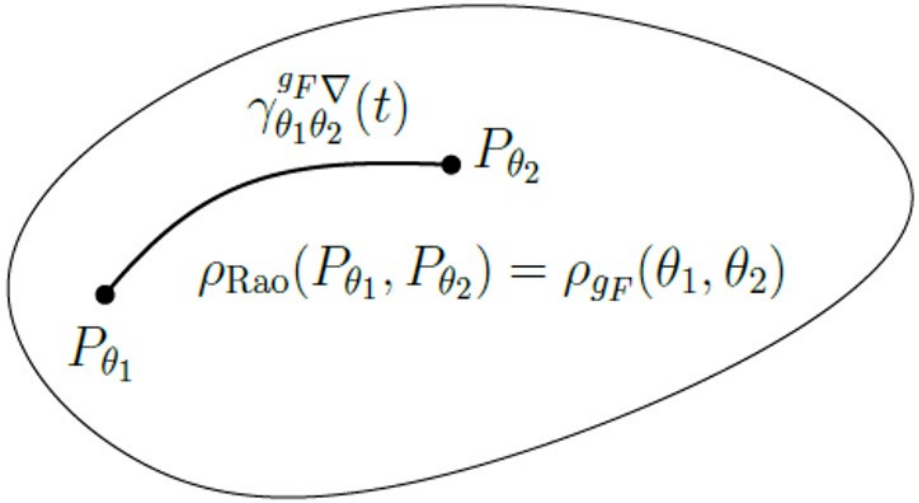
$$L_\theta(\theta) = L_\eta(\eta(\theta))$$

Trouve de très nombreuses applications en optimisation :

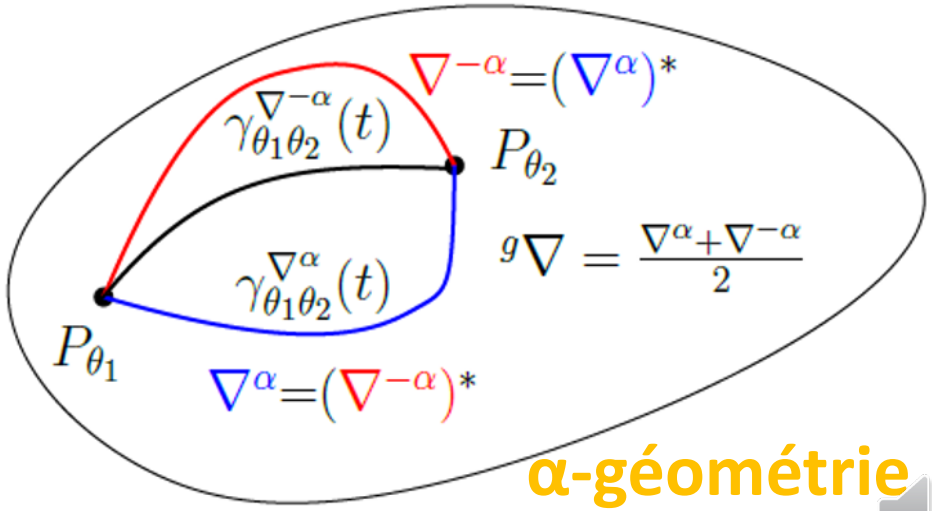
stratégies d'évolutions naturelles (NES), inférence Bayésienne, etc.

La géométrie de l'information en résumé

- **Structures géométriques** pour une famille de lois : le **modèle statistique**
- **Invariance** vis-à-vis du **paramétrage des lois** (θ) et de la **statistique suffisante** (sur l'univers Ω). La distance ne peut croître par une transformation measurable Y=t(X), et reste égale par transformation suffisante
- **Géométrie de Fisher-Rao** avec la distance Riemannienne métrique de Rao
- **α-géométrie duale** (pas nécessairement de divergences associées)
- Interprète le lien étroit entre **estimateur** (maximum de vraisemblance) et **modèle** (maximum d'entropie): **Théorème de Pythagore** et **projections informationelles** dans les **espaces duallement plats**



géométrie de Fisher-Rao



α-géométrie duale

Merci de votre attention

https://franknielsen.github.io/IG/
https://franknielsen.github.io/GSI

Quelques références bibliographiques

- Shun-ichi Amari, ***Information Geometry and Its Applications***, Springer (2016)
- Frank Nielsen, ***The Many Faces of Information Geometry***, Notices of the AMS, 69.1 (2022)
- Frank Nielsen, ***What is an information projection?,*** Notices of the AMS 65.3 (2018)
- Frank Nielsen, **An information-geometric characterization of Chernoff information**, *IEEE Signal Processing Letters* 20.3 (2013): 269-272.
- Vaden Masrani, Rob Brekelmans, Thang Bui, Frank Nielsen, Aram Galstyan, Greg Ver Steeg, Frank Wood, **q-Paths: Generalizing the geometric annealing path using power means**, UAI 2021.
- Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. **Bregman Voronoi diagrams**, *Discrete & Computational Geometry* 44.2 (2010): 281-307.

Gradient naturel et réseaux de neurones artificiels :

- Ke Sun et Frank Nielsen, ***Relative Fisher information and natural gradient for learning large modular models***, ICML (2017)
- Wu Lin, Frank Nielsen, Emtiyaz Kahn, et Mark Schmidt, ***Tractable structured natural-gradient descent using local parameterizations***, ICML (2021)

Remerciements

- À tous mes collaborateurs et plus particulièrement Shun-ichi Amari, Frédéric Barbaresco, Jean-Daniel Boissonnat, Gaëtan Hadjeres, Richard Nock et Ke Sun.

- Sony Computer Science Laboratories Inc. (Tokyo)

- École Polytechnique (Palaiseau)

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

ihp
Institut Henri Poincaré

cnrs

see

International Conference on Bayesian and Maximum Entropy methods in Science and Engineering

41st MaxEnt'22 Conference

July 18-22, IHP, Paris

Important dates

- Deadline for short abstract in 1 or 2 page (MDPI format): **5th April 2022**
- Deadline for conference paper in 8 pages (MDPI format) : **14th June 2022**

TUHH INSTITUTE FOR DATA SCIENCE FOUNDATIONS

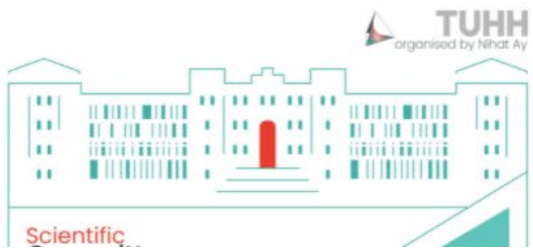
Team Research Teaching Projects Publications Seminars Conferences Location Contact

IG4DS

International Conference on Information Geometry for Data Science

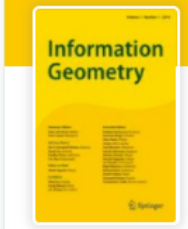
September 19 – 23, 2022, Hamburg, Germany

Home | Confirmed Invited Speakers | Conference Program | Scientific Committee | Location | Organisation | Registration



Traditionally, information geometry has been concerned with the identification of natural geometric structures of statistical models. It has been demonstrated that their use has a great impact on the quality of statistical methods and learning algorithms. One instance of this is given by the natural gradient method, which improves the learning simply by utilising the natural geometry induced by the Fisher-Rao metric. The general geometric perspective of information geometry had already a great influence on machine learning and is expected to further influence the general field of data science.

This conference will bring together scientists from various fields in order to explore the potential of information geometry for the foundations of data science. In addition to invited keynote presentations of leading experts, it will accommodate contributed oral and poster presentations. The submissions of



Information Geometry

Editorial board Aims & scope Journal updates

This journal, as the first to be dedicated to the interdisciplinary field of information geometry:

- Embraces the challenge of uncovering and synthesizing mathematical foundations of information science;
- Offers a platform for intellectual engagements with overlapping interests and diverse backgrounds in mathematical science;
- Balances both theoretical and computational approaches, with ample attention to applications;
- Covers investigations of core concepts defining and studying invariance principles such as the Fisher–Rao metric, dual connection, divergence functions, exponential and mixture geodesics, information projections, and many more areas.

— show all

For authors

Submission guidelines

Ethics & disclosures

Open Access fees and funding

Contact the journal

Submit manuscript

Explore

Online first articles

Volumes and issues

https://www.springer.com/journal/41884