EAI Math Reading Group

Information Geometry

$$g_{ij} = \int_{\Omega} (\partial_i \ln p)(\partial_j \ln p)\, p \, d\omega$$

Motivation

Information Geometry = Differential Geometry applied to statistics

- Statistical Manifold : Manifold of (parametric) distributions
- Fisher Information Metric : Riemannian metric

Applications:

- Cramér-Rao bound
- "distances" between probability distributions
- "shortest-paths" between probability distributions

Outline

1) Recap on Differential Geometry

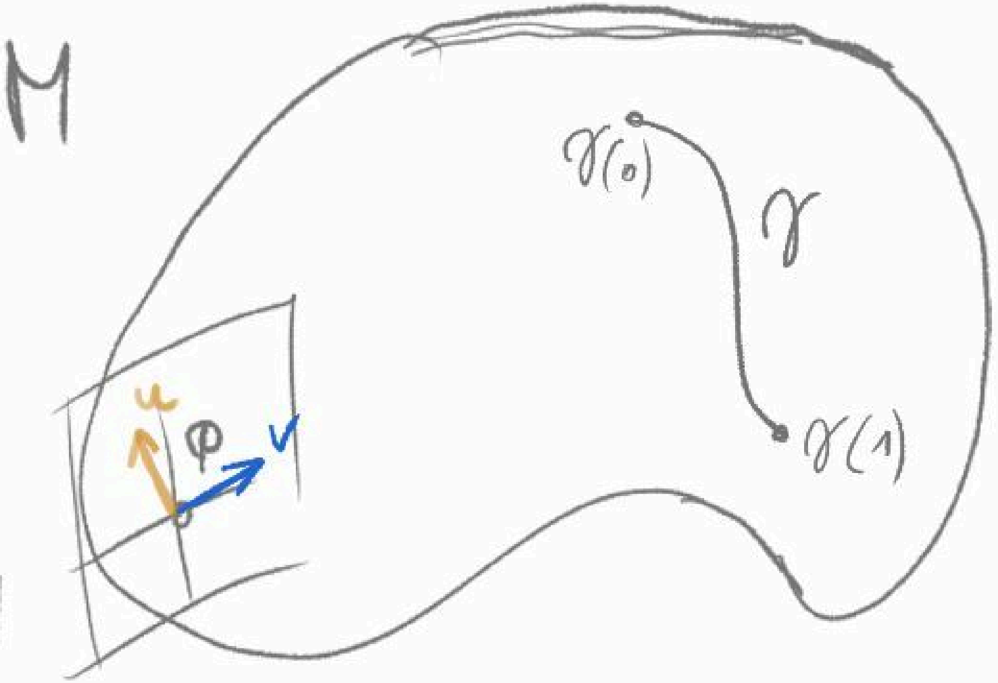
2) The Fisher Information Metric

3) Statistical Manifolds

4) Divergences

① Differential Geometry Recap : Riemannian metric

Manifold M



tangent space T_pM

- inner product

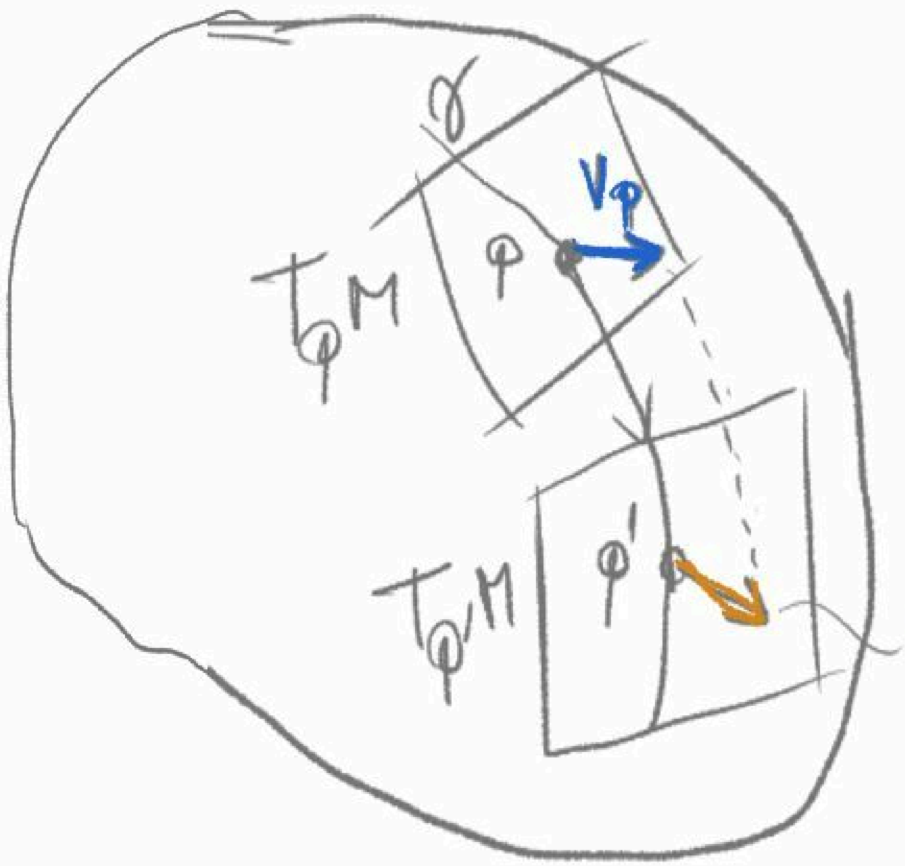
$$\langle u, v \rangle_p = u^i v^j \underbrace{g_{ij}}_{\text{metric}}$$

- length of a path

$$\|\gamma\| = \int_a^b \left\| \frac{d\gamma}{dt} \right\| dt = \int_a^b \sqrt{g_{ij} \dot{\gamma}^i \dot{\gamma}^j} \, dt$$

Riemannian manifold (M, g)

Recap: Affine Connection



$$V_{\varphi'} = \prod_{\gamma(0)\to\gamma(t)}^{\nabla} V \in T_{\varphi'} M$$

$$\approx V_\varphi - dv^i \Gamma_{ij}^{k} \partial_k$$

$$\prod_{\gamma(0)\to\gamma(t)}^{\nabla} \left((\partial_j)_\varphi\right) = (\partial_j)_{\varphi'} - \left(d\theta_i (\Gamma_{ij}^{k})(\partial_k)_{\varphi'}\right)$$

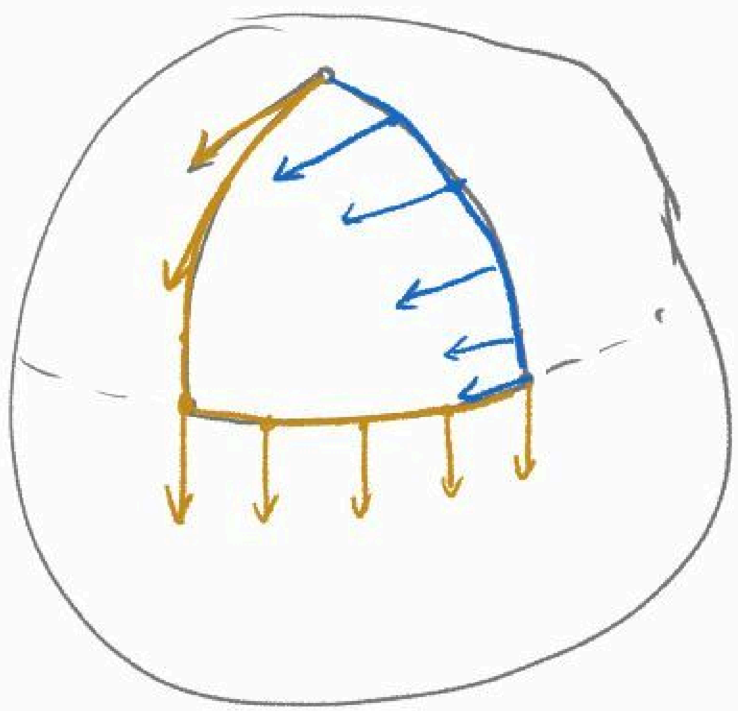
- Affine Connection

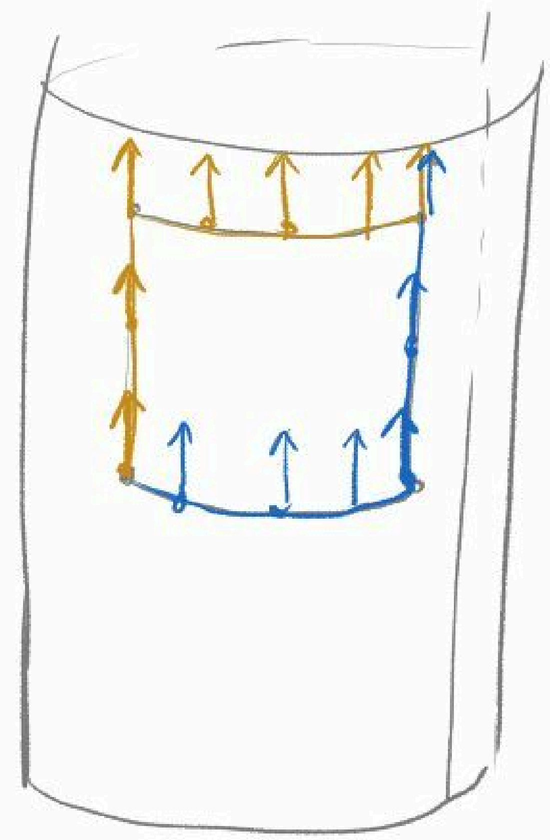
→ covariant derivative

$$(\nabla_X Y)^k = X^i (\nabla_i Y)^k$$
$$= X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma_{ij}^{k} Y^j\right)$$

- parallel transport
- geodesics $\nabla_{\dot\gamma} \dot\gamma = 0$
- curvature, torsion ...

Curvature [the hand wavy version]





constant positive curvature

zero curvature

Metric Connection and Levi-Civita connection

a connection ∇ is compatible with the metric g if

$$Z\langle X,Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle$$

. parallel transport via a metric connection preserves inner products:

$$\left\langle \overset{\nabla}{\underset{\gamma}{\Pi}} u, \overset{\nabla}{\underset{\gamma}{\Pi}} v \right\rangle_{p'} = \langle u, v\rangle_p$$

. There is a unique symmetric (torsion free) connection $\Gamma^k_{ij} = \Gamma^k_{ji}$ compatible with g

called the Levi-Civita connection ${}^{LC}\Gamma^k_{ij}$

Fisher Information Matrix (FIM)

Let $\{p_\theta\}_{\theta \in \Theta}$ be a parametric family of distributions, then the matrix

$$g_{ij}(\theta) = \mathbb{E}_\theta\left[\partial_i l_\theta \, \partial_j l_\theta\right] = \int_\Omega \partial_i l(x;\theta)\, \partial_j l(x;\theta)\, \underbrace{}_{p_\theta(x)} dx$$

with $l(x;\theta) = \log p_\theta(x)$, the log-likelihood, is a symmetric definite positive matrix that can be used as a Riemannian metric

$$p_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Example: (Normal distribution)

$$g_{ij}(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Application: Sufficient Statistics

For some function $Y = F(X)$ w/ $X \sim \varphi(x,\theta)$ we can factor

$$\varphi(x;\theta) = q(F(x);\theta)\underbrace{r(x;\theta)}$$

If $r(x;\theta) = r(x)$ $\rightsquigarrow$ $Y = F(X)$ is a sufficient statistic for X

Theorem

Let $G(\theta)$ be the FIM of $S = \varphi(x;\theta)$ and $G_F(\theta)$ the FIM of $S_F = q(y;\theta)$

then $G_F(\theta) \leq G(\theta)$, i.e. $\underbrace{G(\theta) - G_F(\theta)}$ is definite positive

If F is exhaustive $\Rightarrow$ $G(\theta) - G_F(\theta) = 0$

Proof

$$G_F(\theta)_{ij} = \mathbb{E}_\theta\left[\partial_i \log q(y;\theta)\, \partial_j \log q(y;\theta)\right]$$

- Since $\varphi(x;\theta) = q(F(x);\theta)\, r(x;\theta)$,

$$\partial_i \ell(x;\theta) = \partial_i \log q(F(x);\theta) + \partial_i \log r(x,\theta)$$

- $\partial_i \log q(y;\theta) = \mathbb{E}_\theta\left[\partial_i \ell(X;\theta) \mid Y\right]$:

$$\int_B \partial_i \log q(y;\theta)\, q(y;\theta)\, dy = \int_{F^{-1}(B)} \partial_i \ell(x;\theta)\, \varphi(x;\theta)\, dx$$

- $\Rightarrow \quad \mathbb{E}_\theta\left[\partial_i \log r(X;\theta) \mid F(X)\right] = 0$

$\Rightarrow \partial_i \log r(x;\theta) \perp \phi(F(x))$ for the inner product

$$\langle \phi, \psi \rangle_\theta = \mathbb{E}_\theta \left[\phi(x) \psi(x) \right]$$

$\Rightarrow$ The <u>information loss</u> is given by

$$\underline{\left(\Delta G(\theta) \right)_{ij}} = \mathbb{E}_\theta \left[\partial_i \log r(x;\theta) \, \partial_j \log r(x;\theta) \right] = \mathbb{E}_\theta \left[\underbrace{\mathrm{Cov}_\theta \left[\partial_i \ell, \partial_j \ell \mid Y \right]}_{\geq 0} \right]$$

$\Rightarrow G_F \leq G$ and $G_F = G$ if $\partial_i \log r(x,\theta) = 0$

Cramér-Rao Bound

$$\mathbb{E}[\hat{\theta}] = \theta$$

The variance of an unbiassed estimator $\hat{\theta}$ of θ is at least

$$V_\theta(\hat{\theta}) \geq G(\theta)^{-1}$$

Proof

- $V_\theta(A) = \underbrace{\|(dE[A])_\theta\|^2_\theta}_{\text{norm induced by the Fisher metric.}}$ where A is some random variable

$df \in T^*_\theta$

$df(X) = X(f)$

- For some submanifold $\mathcal{P} \subset \mathcal{P}_X$

$$V_\theta[A] \geq \|(dE[A]|_{\mathcal{P}})_\theta\|^2_\theta$$

so taking $A = \hat{\theta}$

$$V_\theta[\hat{\theta}] \geq \sum_{i,j} (\partial_i E_\theta[\hat{\theta}]\, \partial_j E_\theta[\hat{\theta}])\, G^{-1}_{ij}$$

"Fisher Connection"

Let S be a parametric model. ① Define

$$\left(\Gamma_{ij}^{k}\right)_\theta = \mathbb{E}_\theta\left[\left(\partial_i\partial_j l_\theta + \frac{1}{2}\partial_i l_\theta \partial_j l_\theta\right)(\partial_k l_\theta)\right]$$

- Γ_{ij}^{k} are the Christoffel coefficients of the metric connection associated to the Fisher metric! ($^{LC}\nabla$

- <u>α-connections</u> for $\alpha \in \mathbb{R}$ construct the connection ∇^α from

$$\left(\Gamma_{ij}^{\alpha k}\right)_\theta = \mathbb{E}_\theta\left[\left(\partial_i\partial_j l_\theta + \frac{1-\alpha}{2}\partial_i l_\theta \partial_j l_\theta\right)(\partial_k l_\theta)\right]$$

! $\nabla^0 = {}^{LC}\nabla$

Chentsov Theorem

For a model S, $F(x)$ a sufficient statistic and S_F the induced model, $\partial_i \log p_\theta(x) = \partial_i \log q_\theta(F(x)) \Rightarrow g_{ij}, \Gamma^{(\alpha)k}_{ij}$
coincide

Theorem: Chentsov (1972)

Suppose (g, ∇) invariant for sufficient statistics.

Then there exists $c \in \mathbb{R}_{\geq 0}$ and $\alpha \in \mathbb{R}$ s.t.

cg is the Fisher metric and

$$\nabla = \nabla^{\alpha}$$

Conjugate Connections

Let (M, g, ∇) be a manifold w/ connection ∇, a connection ∇^* is conjugate to ∇ w/ respect to g if

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X^* Z \rangle \quad \text{for } X, Y, Z \in \mathfrak{X}(M)$$

$$\Leftrightarrow \quad X g(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X^* Z)$$

- $(\nabla^*)^* = \nabla$
- Dual Parallel transport preserves the metric:

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

$\rightsquigarrow$ (M, g, ∇, ∇^*) is called a Conjugate Connection Manifold

N.B. $\bar{\nabla} = \frac{\nabla + \nabla^*}{2}$ recovers the Levi-Civita connection

Statistical Manifold

- Amari-Chentsov tensor: $\boxed{C_{ijk} = \Gamma_{ij}^{k} - \overset{*}{\Gamma}_{ij}^{k}}$

  ↳ totally symmetric (0,3) tensor $C_{ijk} = C_{\sigma(i)\sigma(j)\sigma(k)}$ for a permutation

- Statistical manifold (M, g, C)
  ↳ totally symmetric

- α - CCMs: For any pair (∇, ∇^*) of conjugate connections, $\alpha \in \mathbb{R}$, let

$$\begin{cases} \Gamma_{ij}^{\alpha,k} = \Gamma_{ij}^{0k} - \frac{\alpha}{2} C_{ijk} \\ \Gamma_{ij}^{-\alpha,k} = \Gamma_{ij}^{0,k} + \frac{\alpha}{2} C_{ijk} \end{cases}$$

where $\Gamma_{ij}^{0k} = {}^{LC}\Gamma_{ij}^{k}$

$\nabla = \nabla^{1}$, $\nabla^* = \nabla^{-1}$

$\rightsquigarrow$ $(M, g, \nabla^{-\alpha}, \underset{\underset{(\nabla^{-\alpha})^*}{\shortparallel}}{\nabla^{\alpha}})$ is a conjugate connection manifold

Fundamental Theorem of Information Theory

If a torsion-free (symmetric) connection has constant curvature K then so does its conjugate ∇^*

$\Rightarrow$ $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ is ∇^{α}-flat $\Leftrightarrow$ $\nabla^{-\alpha}$-flat

∇-flat $\Leftrightarrow$ ∇^*-flat

$\hookrightarrow \Gamma_{ij}^{k} = 0$

Why do we care?

- Geodesics are easy to compute in flat space (affine curves)

- Adjust α to get a flat space

Divergences

$$D : M \times M \to [0,\infty) \quad , \; D \in C^3$$

1. $D(\theta:\theta') \geq 0$ and $D(\theta:\theta') = 0 \iff \theta = \theta'$

2. $\partial_{i,\cdot} D(\theta:\theta')\big|_{\theta=\theta'} = 0, \; \partial_{\cdot,j} D(\theta:\theta')\big|_{\theta=\theta'} = 0$

3. $-\partial_{\cdot,i}\,\partial_{\cdot,j} D(\theta:\theta')\big|_{\theta'=\theta}$ is positive definite

Example: Kullback-Liebler divergence

$$D(p\,\|\,q) = \int_\Omega \frac{p}{\cancel{q}} \ln \frac{p}{q}\, \cancel{q}\, d\omega$$

$$g_{ij}(\theta) = \int_\Omega (\partial_i \ln p)(\partial_j \ln p_\theta)\, p_\theta \, d\omega$$

Fisher Matrix !

$\mathcal{D}$ivergence to manifold

given a divergence $\mathcal{D}$, let

$$ {}^{\mathcal{D}}g = -\partial_{i,j}\, \mathcal{D}(\theta : \theta')\big|_{\theta=\theta'} $$

$$ \Gamma_{ijk} = -\partial_{ij,k}\, \mathcal{D}(\theta : \theta')\big|_{\theta=\theta'} $$

$$ \Gamma^{*}_{ijk} = -\partial_{k,ij}\, \mathcal{D}(\theta : \theta')\big|_{\theta=\theta'} $$

then $(M, {}^{\mathcal{D}}g, {}^{\mathcal{D}}\nabla, {}^{\mathcal{D}}\nabla^{*})$ is a ecm