- Remarks on the Nielsen review paper
  - arXiv: 1808.08271

- Information geometry (IG) includes non-Riemannian manifolds
  - but IG historically started w/ the Riemannian modeling $(\mathcal{P}, \mathcal{P}_g)$ of a parametric family of prob. dist. $\mathcal{P}$ by letting the metric tensor be the Fisher information matrix.
- Can define an information manifold from a divergence.

- Outline of the Nielson review paper. (1.2. Outline)
  - Ch. 2 : Riemannian geometry
    - manifold (M, g, ∇)
    - Riemannian manifolds (M, g)
      - unique torsion-free Levi-Civita ∇ from g
  - Ch 3 : Information geometry
    - conjugate connection manifolds (M, g, ∇, ∇^*)
    - statistical manifolds (M, g, C)
    - a family of information manifolds $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$
    - conjugate connections from divergences
    - dually flat manifolds from Bregman divergences
    - exponential connection ${}^e\nabla$ & mixture connection ${}^m\nabla$ coupled to the Fisher information matrix for parametric family of prob. models.
  - Ch 4
    - examples

- a smooth D-dim manifold M
  - a topological space that locally looks like $\mathbb{R}^D$
- points, vectors, functions, diff. ops., live on M and coord-free but can be expressed in any <u>local</u> coord. of $\mathcal{A} = \{(U_i, x_i)\}_i$
- At each $p \in M$, $\exists$ tangent space T_p that locally approximates M as a vector space.
- On M, can define two <u>indep.</u> structures
  - a metric tensor field g
  - an affine connection ∇ (what is an example of non-affine ∇?)
- g defines an inner prod. of vectors on T_p
- ∇ is a diff. op. that defines
  - the covariant derivative $\nabla_X Y$ for vector fields X, Y
  - the parallel transport Π_c^∇ along a smooth curve c
  - ∇-geodesic γ_∇
  - curvature
  - torsion

- Christoffel symbols $\Gamma_{ij}^{k}(p)$ define the unique affine connection ∇ that satisfies

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^{k}\,\partial_k \longrightarrow (\nabla_{\partial_i}\partial_j)^k = \Gamma_{ij}^{k}$$

  - $(\nabla_X Y)^k = X^i(\nabla_i Y)^k = X^i(\partial_i Y^k + \Gamma_{ij}^{k} Y^j)$
  - Γ is <u>not a tensor field.</u>

- ∇ defines how to transport vectors at T_p to T_q along $c(t)$, $c(t=0)=p$, $c(t=1)=q$., called parallel transport Π.

$$\forall v \in T_p,\ \forall t \in [0,1],\ v_{c(t)} = \overset{\nabla}{\underset{c(0)\to c(t)}{\Pi}} v \in T_{c(t)}$$

- geodesic curve $\gamma(t)$: $\nabla_{\dot\gamma}\dot\gamma = 0$
- curvature tensor field R^i_{jkl} (Riemannian)

$$R(X,Y)Z = \nabla_X\nabla_Y X - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z$$

measures the deviation of Z from itself after a parallel transport along an infinitesimal rectangle formed by X & Y

- various flatness according to which R we use.
  - Ricci curvature $R_{ab} = R^c{}_{acb} = K g_{ab}$, K: Gaussian curvature
- torsion $T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]$
  - M is torsion-free $\longleftrightarrow$ ∇ is symmetric
  - how tangent spaces twist when parallel transported.

- ∇ is metric-compatible when
$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$$
  - such ∇ preserves the metric under Π^∇:
$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla} v \right\rangle, \quad \forall t$$
- Levi-Civita connection ${}^{LC}\nabla$
  - A <u>unique</u>, <u>torsion-free</u>, <u>affine connection</u>, <u>compatible with the metric</u>

$${}^{LC}\Gamma^k_{ij} = \frac{1}{2} g^{kl} (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij})$$

$$\Leftrightarrow 2g(\nabla_X Y, Z) =$$
$$= X(g(Y,Z)) + Y(g(X,Z)) - Z(g(X,Y))$$
$$+ g([X,Y],Z) - g([X,Z],Y) - g([Y,Z],X)$$

- ${}^{LC}\nabla$ determines the local structure of M, but not the global topology
  - cone and cylinder have the same locally flat Euclidean metric.

- Conjugate connection manifolds (M, g, ∇, ∇^*)
  - conjugate connections w.r.t. g:

$$\underbrace{X g(Y, Z)}_{\text{directional deriv. of a scalar w.r.t. } X} = \underbrace{g(\nabla_X Y, Z) + g(Y, \nabla_X^* Z)}_{\text{scalar}}$$

  - cf. Riemannian manifold (M, g), $\nabla = \nabla^* = {}^{LC}\nabla$

    information geometric manifolds does not induce a distance (why?)
  - (∇, ∇^*) preserves g iff

$$\forall t \in [0,1], \left\langle \prod^{\nabla}_{c(0)\to c(t)} u, \prod^{\nabla^*}_{c(0)\to c(t)} v \right\rangle = \langle u, v \rangle_{c(0)}$$

  - for (M, g, ∇), $\exists!\ \nabla^*$
  - $$\bar{\nabla} = \frac{\nabla + \nabla^*}{2} = {}^{LC}\nabla$$

- Statistical manifolds (M, g, C)
  - Amari–Chentsov tensor

    $C_{ijk} = \Gamma_{ij}^{k} - \Gamma^{*k}_{ij} \Longleftrightarrow C(X,Y,Z) = \langle \nabla_X Y - \nabla_X^* Y, Z \rangle$
    - totally symmetric $(0,3)$-tensor

- A family of conjugate connection manifolds $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$

  $\Longleftrightarrow$ family of information <u>α-manifolds</u>
  - α-connections $(\nabla^{-\alpha}, \nabla^{\alpha})$, $\alpha \in \mathbb{R}$
    - $\nabla^{-1} = \nabla^*$, $\nabla^0 = \bar{\nabla} = {}^{LC}\nabla$, $\nabla^1 = \nabla$

- ∇ is κ-curved $\Longleftrightarrow$ ∇^* is κ-curved
  - $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ is ∇^{α}-flat $\Longleftrightarrow$ $\nabla^{-\alpha}$-flat
  - 〃 〃 ∇-flat $\Longleftrightarrow$ ∇^*-flat

- Divergence D on a manifold M

  $D: M \times M \to [0, \infty]$

  - smooth "distance", possibly symmetric.
  - example : squared Euclidean distance
  - not : Euclidean distance, Riemannian metric distnce D_ρ.

    $$D_\rho(p, q) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)}\, dt$$

  - D defines $(M, D) = (M, {}^D g, {}^D\nabla, {}^D\nabla^*)$
    - a conjugate connection manifold $(M, {}^D g, {}^D\nabla, {}^{D^*}\nabla)$
    - a statistical manifold $(M, {}^D g, {}^D C)$
    - a one-param family of conjugat conn. manifolds

      $\{(M, {}^D g, {}^D C^\alpha) = (M, {}^D g, {}^D\nabla^{-\alpha}, ({}^D\nabla^{-\alpha})^* = {}^D\nabla^\alpha)\}_{\alpha \in \mathbb{R}}$

- Dually flat manifolds, Bregman geometry,
  - it is a flat manifold, therefore can use various tools of flat geometry like Pythagorean theorems, ustering, ...
  - For a strictly convex smooth function $F(\theta)$ called a potential function, Bregman divergence (parameter divergence) is
    $$B_F(\theta : \theta') = F(\theta) - F(\theta') - (\theta - \theta')^T \cdot \nabla F(\theta')$$
  - induced information-geometric structure is $(M, {}^{B_F}g, {}^{B_F}c)$
  - example of F
    - for an exponential family $\mathcal{E}$,
      $$\mathcal{E} = \left\{ p_\theta(x) = \exp\left(\sum_{i=1}^{D} t_i(x)\theta_i - F(\theta) + k(x) \right) \right\}$$

- Divergence, α-connection, Fisher information matrix

- Invariant standard f-divergences yield
  - the Fisher information matrix, and
  - the α-connection

$$I_f[p(x;\theta) : p(x;\theta+d\theta)]$$

$$= \int p(x;\theta) \cdot f\left(\frac{p(x;\theta+d\theta)}{p(x;\theta)}\right) d\mu(x)$$

$$= \frac{1}{2} \cdot {}_F g_{ij}(\theta)\, d\theta^i d\theta^j$$

- Expected α-manifolds of a family of parametric prob. dist. $(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{\alpha})$
  - $\underline{\mathcal{P} = \{ p_\theta(x) \}}_{\theta \in \Theta}$, $\dim(\Theta) = D$
  - $l(\theta; x) = \log p_\theta(x)$: log-likelihood
  - the Fisher information matrix (FIM)
    - $D \times D$ matrix
    - $\underline{{}_{\mathcal{P}}I(\theta)} = E[\partial_i l \, \partial_j l \mid \theta]$
    - invariant by reparametrization of the sample space $\mathcal{X}$
    - covariant 〃 〃 of the param space Θ
  - the Fisher information metric tensor
    - built from the FIM as

$$\underline{{}_{\mathcal{P}}g(u,v)} = [{}_{\mathcal{P}}I(\theta)]_{ij} \cdot (u_\theta)_i (v_\theta)_j$$

  - expected exponential connection ${}^{e}_{\mathcal{P}}\nabla$

$${}^{e}_{\mathcal{P}}\nabla = E[(\partial_i \partial_j l)(\partial_k l) \mid \theta]$$

  - expected mixture connection ${}^{m}_{\mathcal{P}}\nabla$

$${}^{m}_{\mathcal{P}}\nabla = E[(\partial_i \partial_j l + \partial_i l \, \partial_j l)(\partial_k l) \mid \theta]$$

  - $(\mathcal{P}, {}_{\mathcal{P}}g, {}^{m}_{\mathcal{P}}\nabla, {}^{e}_{\mathcal{P}}\nabla) \longrightarrow \{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\}_{\alpha \in \mathbb{R}}$

$${}_{\mathcal{P}}\Gamma^{\alpha}_{ij,k} = E\left[\left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \, \partial_j l\right)(\partial_k l) \,\middle|\, \theta\right]$$

$${}_{\mathcal{P}}T^{-1}_{ij,k} = ({}^{m}_{\mathcal{P}}\nabla)_{ij,k}, \quad {}_{\mathcal{P}}T^{+1}_{ij,k} = ({}^{e}_{\mathcal{P}}\nabla)_{ij,k}$$

  - ${}^{e}_{\mu}T = {}^{m}_{\mu}T = {}^{e}_{\varepsilon}T = {}^{m}_{\varepsilon}T = 0$

- f-divergence
  - for a convex f,

    $$I_f(\theta : \theta') = \sum_{i=1}^{D} \theta_i f(\theta_i' / \theta_i) \quad , \quad f(1) = 0$$

  - Wikipedia:

    $$D_f(P \| Q) = \int_\Omega f\left(\frac{dP}{dQ}\right) dQ \quad , \quad f(1) = 0$$

    when $dP = p\,d\mu$ & $dQ = q\,d\mu$, then

    $$D_f(P \| Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x)$$

  - examples
    - family of α-divergences including D_{KL} ($\alpha \to 1$)

      $$f(u) = \frac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) \quad , \quad \lim_{\alpha \to 1} f(u) = u \log u$$

    - Jenson-Shannon D_{JS}, $f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$

- Invariant standard f-divergences yield
  - the Fisher information matrix, and
  - the α-connection

$$D_f[p(x;\theta):p(x;\theta+d\theta)]$$

$$= \int p(x;\theta)\cdot f\left(\frac{p(x;\theta+d\theta)}{p(x;\theta)}\right) d\mu(x)$$

$$\partial_i l = \partial_i \log p(x;\theta) \quad \Big| \; p(x;\theta+d\theta) = p(x;\theta)$$

$$= \partial_i p \cdot d\theta_i / p(x;\theta) \quad \Big| \; + \partial_i p \cdot d\theta_i$$

$$f(p(x;\theta+d\theta)/p(x;\theta))$$

$$= f(1 + \sum_i \partial_i p \cdot d\theta_i / p(x;\theta))$$

$$\cong \underbrace{f(1)}_{=0} + \underbrace{f'(1)}_{=0}\cdot\Delta + \underbrace{f''(1)}_{=1}\cdot\frac{1}{2}\Delta^2$$

need to check these by exactly considering $O(d\theta^2)$ terms.

$$\Rightarrow D_f[p(x;\theta):p(x;\theta+d\theta)]$$

$$= \frac{1}{2}\, {}_F g_{ij}(\theta)\, d\theta^i d\theta^j$$

$$\partial^2 \log f = \partial^2 f/f - (\partial f/f)^2 = \partial^2 f/f - (\partial \log f)^2$$

$$E[\partial^2 f/f \mid \theta] = \partial^2 \int f(x;\theta)\,dx = 0$$

$$\Rightarrow E[\partial^2 \log f] = -E[(\partial \log f)^2 \mid \theta]$$

- Fisher - Rao expected Riemannian manifold $(\mathcal{P}, {}^{p}g)$
  (Fisher - Riemannian manifolds)
  - family of categorical dist. : spherical
  - family of bivariate location-scale families: hyperbolic
  - family of location families : Euclidean

- Examples of dually-flat structures
    - dually flat exponential manifold, induced by reverse KL, on an exponential family
    - dually flat mixture manifold, induced by KL, on a mixture family