Recursive estimation in Riemannian manifolds

Salem Said 2019 – École d'été de Peyresq

CNRS — Université de Bordeaux

You should become able to read this :

Zhou & Said (2018) : fast asymptotically-efficient recursive estimation in a Riemannian manifold (https ://arxiv.org/abs/1805.06811)

Nothing is more practical than a good theory !!

Plan

What is a manifold ?

◇ Riemann (1854) : any space which one can describe by varying n real parameters

{attitude of a solid body in space}

{shape of a deformable elastic body}

{color}

◇ Abstract manifolds (1920s-30s) : the cartographer's definition

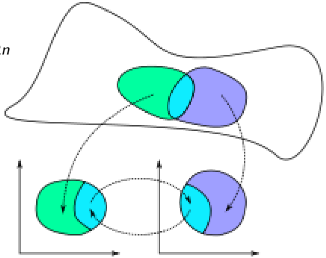
$$\{(z,w) \in \mathbb{C} \times \mathbb{C} | w^2 - z^2 + 4 = 0\} \text{ is a cylindre !!}$$

a manifold is a space with at atlas which is a set of compatible local charts

◇ Whitney's theorem (1944) : all manifolds are concrete

–any smooth n-dimensional manifold can be embedded into $\mathbb{R}^{2n}$

– the embedding is difficult, but we know it always exists



What is a Riemannian manifold ?

◇ Riemannian metric : length is measured by a quadratic form

---

think of an embedded manifold $M \subset \mathbb{R}^N$

$$L(\gamma) = \int_0^1 \langle \dot\gamma(t), \dot\gamma(t) \rangle^{\frac{1}{2}} \, dt \qquad \text{curve } \gamma : [0.1] \to M$$

⇝ reparameterisation-invariant definition of length $L(\gamma \circ \phi) = L(\gamma)$

◇ Riemannian distance :

$$d(x, y) = \inf\{L(\gamma)\,;\ \gamma(0) = x \text{ and } \gamma(1) = y\}$$

◇ Geodesics : locally length-minimising curves

$$L(\gamma|[t, t+\epsilon]) = d(\gamma(t), \gamma(t+\epsilon)) \qquad \text{for each } t \in (0, 1)$$

a geodesic may fail to be globally minimising!!

◇ there exists an infinite choice of Riemannian metrics on a given manifold

◇ Curvature is a more relevant quantity

Gauss-Bonnet for surfaces : $\displaystyle \frac{1}{4\pi} \int_M R \, d\mathrm{vol} = 2 - 2g$

What is a Riemannian manifold ?

◇ Riemannian metric : length is measured by a quadratic form

---

think of an embedded manifold $M \subset \mathbb{R}^N$

$$L(\gamma) = \int_0^1 \langle \dot\gamma(t), \dot\gamma(t) \rangle^{\frac{1}{2}} \, dt \qquad \text{curve } \gamma : [0.1] \to M$$

⇝ reparameterisation-invariant definition of length $\qquad L(\gamma \circ \phi) = L(\gamma)$

◇ Riemannian distance :

$$d(x, y) = \inf \{ L(\gamma) \, ; \, \gamma(0) = x \text{ and } \gamma(1) = y \}$$

◇ Geodesics : locally length-minimising curves

ϵ essentially depends on current point $\gamma(t)$

$$L(\gamma | [t, t+\epsilon]) = d(\gamma(t), \gamma(t+\epsilon)) \qquad \text{for each } t \in (0,1)$$

a geodesic may fail to be globally minimising!!

◇ there exists an infinite choice of Riemannian metrics on a given manifold

◇ Curvature is a more relevant quantity

Gauss-Bonnet for surfaces : $\dfrac{1}{4\pi} \displaystyle\int_M R \, d\mathrm{vol} = 2 - 2g$

Curvature and distance function

◇ Geodesic spherical coordinates : fix some point $x \in M$

$$(r, \theta) \longmapsto \gamma_\theta(r) \qquad \text{for } \; r \in (0, \epsilon_x) \;,\; \theta \in S^1$$

◇ The metric :

$$\underbrace{\langle v, v\rangle \;=\; v_r^2 + f^2(r, \theta)\, v_\theta^2}_{\text{scalar product}} \qquad \underbrace{ds^2 = dr^2 + f^2(r, \theta)\, d\theta^2}_{\text{length element}}$$

◇ Jacobi equation : second order linear ode

$$f_{rr} + K f = 0 \qquad K(r, \theta) \;\; \text{sectional curvature}$$

◇ Constant curvature :

$$\underbrace{f(r, \theta) \;=\; k^{-1}\, \sin(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{f(r, \theta) = k^{-1}\, \sinh(kr)}_{\text{negative curvature } -k^2}$$

◇ Distance function (locally!!) : $y = \gamma_\theta(r) \implies d(x, y) = r$

◇ Hessian of distance function :

$$\nabla^2 r \simeq \begin{pmatrix} 0 & 0 \\ 0 & f_r/f \end{pmatrix}$$

Curvature and distance function

◇ Geodesic spherical coordinates : fix some point $x \in M$

γ_θ unit speed geodesic in direction θ

$$(r, \theta) \longmapsto \gamma_\theta(r) \qquad \text{for} \quad r \in (0, \epsilon_x)\ ,\ \theta \in S^1$$

◇ The metric :

$$\underbrace{\langle v, v\rangle \ = \ v_r^2 + f^2(r, \theta)\, v_\theta^2}_{\text{scalar product}} \qquad \underbrace{ds^2 = dr^2 + f^2(r, \theta)\, d\theta^2}_{\text{length element}}$$

◇ Jacobi equation : second order linear ode

$$f_{rr} + K f = 0 \qquad K(r, \theta) \ \text{sectional curvature}$$

◇ Constant curvature :

$$\underbrace{f(r, \theta) \ = \ k^{-1} \, \sin(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{f(r, \theta) = k^{-1} \, \sinh(kr)}_{\text{negative curvature } -k^2}$$

◇ Distance function (locally!!) : $y = \gamma_\theta(r) \implies d(x, y) = r$

◇ Hessian of distance function :

$$\nabla^2 r \simeq \begin{pmatrix} 0 & 0 \\ 0 & f_r/f \end{pmatrix}$$

Curvature and distance function

◇ Geodesic spherical coordinates : fix some point $x \in M$

γ_θ unit speed geodesic in direction θ

$$(r, \theta) \longmapsto \gamma_\theta(r) \qquad \text{for } r \in (0, \epsilon_x)\ ,\ \theta \in S^1$$

◇ The metric :

$$\underbrace{\langle v, v\rangle = v_r^2 + f^2(r,\theta)\, v_\theta^2}_{\text{scalar product}} \qquad \underbrace{ds^2 = dr^2 + f^2(r,\theta)\, d\theta^2}_{\text{length element}}$$

◇ Jacobi equation : second order linear ode

K determines M up to coverings

$$f_{rr} + K f = 0 \qquad K(r,\theta) \text{ sectional curvature}$$

◇ Constant curvature :

$$\underbrace{f(r,\theta) = k^{-1}\sin(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{f(r,\theta) = k^{-1}\sinh(kr)}_{\text{negative curvature } -k^2}$$

◇ Distance function (locally!!) :

$$y = \gamma_\theta(r) \implies d(x,y) = r$$

◇ Hessian of distance function :

$$\nabla^2 r \simeq \begin{pmatrix} 0 & 0 \\ 0 & f_r/f \end{pmatrix}$$

Curvature and distance function

◇ Geodesic spherical coordinates :

fix some point $x \in M$

γ_θ unit speed geodesic in direction θ

$$(r,\theta) \longmapsto \gamma_\theta(r) \qquad \text{for} \quad r \in (0, \epsilon_x)\ ,\ \theta \in S^1$$

◇ The metric :

$$\underbrace{\langle v, v\rangle = v_r^2 + f^2(r,\theta)\, v_\theta^2}_{\text{scalar product}} \qquad \underbrace{ds^2 = dr^2 + f^2(r,\theta)\, d\theta^2}_{\text{length element}}$$

◇ Jacobi equation : second order linear ode

K determines M up to coverings

$$f_{rr} + K f = 0 \qquad K(r,\theta) \ \text{sectional curvature}$$

◇ Constant curvature :

$f(r,\theta) \sim r$ for small r

$$\underbrace{f(r,\theta) = k^{-1}\sin(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{f(r,\theta) = k^{-1}\sinh(kr)}_{\text{negative curvature } -k^2}$$

◇ Distance function (locally!!) :

$$y = \gamma_\theta(r) \implies d(x,y) = r$$

◇ Hessian of distance function :

$$\nabla^2 r \simeq \begin{pmatrix} 0 & 0 \\ 0 & f_r/f \end{pmatrix}$$

Curvature and distance function

◇ Geodesic spherical coordinates : fix some point $x \in M$

γ_θ unit speed geodesic in direction θ

$$(r,\theta) \longmapsto \gamma_\theta(r) \qquad \text{for} \quad r \in (0,\epsilon_x)\ ,\ \theta \in S^1$$

◇ The metric :

$$\underbrace{\langle v,v\rangle = v_r^2 + f^2(r,\theta)\, v_\theta^2}_{\text{scalar product}} \qquad \underbrace{ds^2 = dr^2 + f^2(r,\theta)\, d\theta^2}_{\text{length element}}$$

◇ Jacobi equation : second order linear ode

K determines M up to coverings

$$f_{rr} + K f = 0 \qquad K(r,\theta) \ \text{sectional curvature}$$

◇ Constant curvature :

$f(r,\theta) \sim r$ for small r

$$\underbrace{f(r,\theta) = k^{-1}\sin(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{f(r,\theta) = k^{-1}\sinh(kr)}_{\text{negative curvature } -k^2}$$

◇ Distance function (locally!!) : $y = \gamma_\theta(r) \implies d(x,y) = r$

◇ Hessian of distance function :

$$\nabla^2 r \simeq \begin{pmatrix} 0 & 0 \\ 0 & f_r/f \end{pmatrix}$$

f_r/f may become negative or diverge !!

Hessian comparison

◇ Constant curvature : "curvature of a sphere" $S = f_r/f$

$$\underbrace{S(r,\theta) = k\cot(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{S(r,\theta) = k\coth(kr)}_{\text{negative curvature } -k^2}$$

⇝ distance can fail to be convex or smooth!!

◇ Comparison :

$$\alpha \le K(r,\theta) \le \beta \implies \underbrace{S_\beta(r) \le S(r,\theta) \le S_\alpha(r)}_{\text{note reverse order}}$$

⇝ some local estimates

$$\frac{\pi}{2\sqrt{\beta}} \le r \le \frac{\pi}{2\sqrt{\alpha}} \qquad \frac{\pi}{\sqrt{\beta}} \le r \le \frac{\pi}{\sqrt{\alpha}}$$

convexity is lost conjugate points (foci)

◇ Hessian of squared distance :

$$e_x(y) = d^2(x,y) \qquad \nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2rS \end{pmatrix}$$

Hessian comparison

◇ Constant curvature : "curvature of a sphere" $S = f_r/f$

$S \sim r^{-1}$ *for small* r

$$\underbrace{S(r,\theta) = k\cot(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{S(r,\theta) = k\coth(kr)}_{\text{negative curvature } -k^2}$$

⇝ distance can fail to be convex or smooth!!

◇ Comparison :

$$\alpha \le K(r,\theta) \le \beta \implies \underbrace{S_\beta(r) \le S(r,\theta) \le S_\alpha(r)}_{\text{note reverse order}}$$

⇝ some local estimates

$$\frac{\pi}{2\sqrt{\beta}} \le r \le \frac{\pi}{2\sqrt{\alpha}} \qquad \frac{\pi}{\sqrt{\beta}} \le r \le \frac{\pi}{\sqrt{\alpha}}$$

convexity is lost   conjugate points (foci)

◇ Hessian of squared distance :

$$e_x(y) = d^2(x,y) \qquad \nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2rS \end{pmatrix}$$

Hessian comparison

◇ Constant curvature : "curvature of a sphere" $S = f_r/f$

$S \sim r^{-1}$ for small r

$$\underbrace{S(r,\theta) = k\cot(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{S(r,\theta) = k\coth(kr)}_{\text{negative curvature } -k^2}$$

$\rightsquigarrow$ distance can fail to be convex or smooth!!

◇ Comparison :

$$\alpha \le K(r,\theta) \le \beta \implies \underbrace{S_\beta(r) \le S(r,\theta) \le S_\alpha(r)}_{\text{note reverse order}}$$

$\rightsquigarrow$ some local estimates

non-positive curvature
is well behaved

$$\frac{\pi}{2\sqrt{\beta}} \le r \le \frac{\pi}{2\sqrt{\alpha}} \qquad \frac{\pi}{\sqrt{\beta}} \le r \le \frac{\pi}{\sqrt{\alpha}}$$

convexity is lost conjugate points (foci)

◇ Hessian of squared distance :

$$e_x(y) = d^2(x,y) \qquad \nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2rS \end{pmatrix}$$

Hessian comparison

◇ Constant curvature :   "curvature of a sphere" $S = f_r/f$

$S \sim r^{-1}$ for small r

$$\underbrace{S(r,\theta) \;=\; k\,\cot(kr)}_{\text{positive curvature } k^2} \qquad \underbrace{S(r,\theta) = k\,\coth(kr)}_{\text{negative curvature } -k^2}$$

⇝ distance can fail to be convex or smooth!!

◇ Comparison :

$$\alpha \le K(r,\theta) \le \beta \implies \underbrace{S_\beta(r) \le S(r,\theta) \le S_\alpha(r)}_{\text{note reverse order}}$$

⇝ some local estimates

non-positive curvature
is well behaved

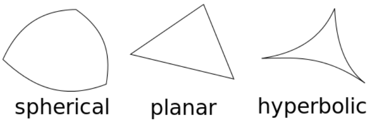
$$\frac{\pi}{2\sqrt{\beta}} \le r \le \frac{\pi}{2\sqrt{\alpha}} \qquad \frac{\pi}{\sqrt{\beta}} \le r \le \frac{\pi}{\sqrt{\alpha}}$$

convexity is lost   conjugate points (foci)

◇ Hessian of squared distance :

$$\nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 2(rS-1) \end{pmatrix}$$

Triangle comparison



◇ Geodesic triangle $\Delta = xyz$:

$$(zy)^2 = \underbrace{\left[(zx)^2 - 2(zx)(xy)\cos\angle zxy + (xy)^2\right]}_{\text{planar triangle}} + \mathcal{E}(\Delta)$$

◇ Error estimate :

$$(xy)^2(DS_\beta(D) - 1) \leq \mathcal{E}(\Delta) \leq (xy)^2(DS_\alpha(D) - 1)$$

◇ Sign of K :

$$K \geq 0 \Rightarrow \alpha = 0 \Rightarrow \mathcal{E}(\Delta) \leq 0 \text{ so } (zy) \leq (zy)_{\text{plane}}$$

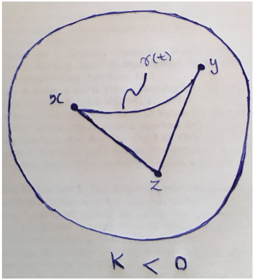
$$K \leq 0 \Rightarrow \beta = 0 \Rightarrow \mathcal{E}(\Delta) \geq 0 \text{ so } (zy) \geq (zy)_{\text{plane}}$$

◇ Other comparisons :

metric, area, sum of angles, ...

$$\underbrace{f_\beta(r) \leq f(r,\theta) \leq f_\alpha(r)}_{f = 0 \text{ before } f_\alpha \text{ and after } f_\beta}$$

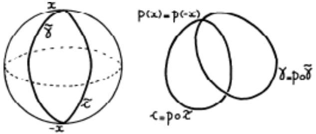
second-order Taylor of
$f(\gamma(t)) = d^2(z, \gamma(t))$



Cut and conjugate locus

$$\mathrm{Cut}(x) = \{\gamma_\theta(r_c) \mid \gamma_\theta \text{ not minimising after } r_c\}$$

$$\mathrm{Conj}(x) = \{r(\theta) \mid f(r,\theta) = 0 \text{ for the first time}\}$$



◇ Why does a geodesic fail to minimise ?

— Contains conjugate points : not a local minimum
— It is not unique (broken geodesics don't minimise)

$$y \in \mathrm{Cut}(x) \iff y \in \mathrm{Conj}(x) \text{ or } \underbrace{y = c(1) = \gamma(1)}_{\text{typical case}}$$

◇ Cut locus and topology

true in any dimension

Cut locus is a negligeable set

$$M = D_x \cup \mathrm{Cut}(x) \qquad \mathrm{Cut}(x) \text{ deformation retract of } M - \{x\}$$

◇ Injectivity radius

$$i(x) = \inf_{y\ \mathrm{Cut}(x)} d(x,y) = d(x, \mathrm{Cut}(x))$$

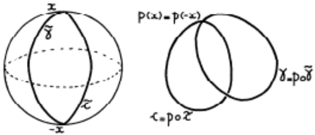
◇ Klingenberg

$$i(x) \geq \min\left\{\frac{\pi}{\sqrt{\beta}}, \frac{\ell}{2}\right\}$$

Cut and conjugate locus

$$\mathrm{Cut}(x) = \{\gamma_\theta(r_c) \mid \gamma_\theta \text{ not minimising after } r_c\}$$

$$\mathrm{Conj}(x) = \{r(\theta) \mid f(r,\theta) = 0 \text{ for the first time}\}$$



◇ Why does a geodesic fail to minimise ?

— Contains conjugate points : not a local minimum
— It is not unique (broken geodesics don't minimise)

$$y \in \mathrm{Cut}(x) \;\Leftrightarrow\; y \in \mathrm{Conj}(x) \text{ or } \underbrace{y = c(1) = \gamma(1)}_{\text{typical case}}$$

◇ Cut locus and topology

true in any dimension

Cut locus is a negligeable set

$$M = D_x \cup \mathrm{Cut}(x) \qquad\qquad \mathrm{Cut}(x) \text{ deformation retract of } M - \{x\}$$

◇ Injectivity radius

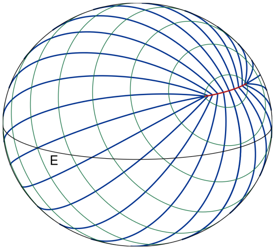
$$i(x) = \inf_{y\ \mathrm{Cut}(x)} d(x,y) = d(x, \mathrm{Cut}(x))$$

◇ Klingenberg

$$i(x) \geq \min\left\{\frac{\pi}{\sqrt{\beta}}, \frac{\ell}{2}\right\}$$

ℓ length of shortest loop through x

Cut and conjugate locus



$$\mathrm{Cut}(x) = \{\gamma_\theta(r_c) \mid \gamma_\theta \text{ not minimising after } r_c\}$$

$$\mathrm{Conj}(x) = \{r(\theta) \mid f(r,\theta) = 0 \text{ for the first time}\}$$

◇ Why does a geodesic fail to minimise ?

— Contains conjugate points : not a local minimum
— It is not unique (broken geodesics don't minimise)

$$y \in \mathrm{Cut}(x) \Leftrightarrow y \in \mathrm{Conj}(x) \text{ or } \underbrace{y = c(1) = \gamma(1)}_{\text{typical case}}$$

◇ Cut locus and topology

true in any dimension

Cut locus is a negligeable set

$$M = D_x \cup \mathrm{Cut}(x) \qquad \mathrm{Cut}(x) \text{ deformation retract of } M - \{x\}$$

◇ Injectivity radius

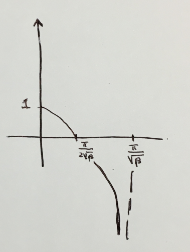
$$i(x) = \inf_{y\ \mathrm{Cut}(x)} d(x,y) = d(x, \mathrm{Cut}(x))$$

◇ Klingenberg

$$i(x) \geq \min\left\{\frac{\pi}{\sqrt{\beta}}, \frac{\ell}{2}\right\}$$

ℓ length of shortest loop through x

Convexity of squared distance



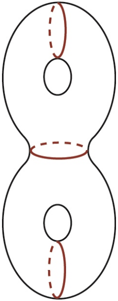
◇ Recall the Hessian :

$$\nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2rS \end{pmatrix} \geq 2\,\beta r \cot(\beta r) \qquad \text{(positive curvature)}$$

$$\geq 2 \qquad \text{(non-positive curvature)}$$

◇ But is it convex ?

The problem comes from closed geodesics

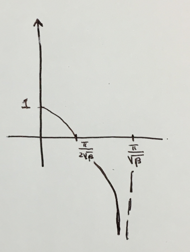


◇ Convexity radius : when is a geodesic ball convex ?

$$B(x, R) \text{ is convex if } \quad R \leq \frac{i(M)}{2} \;\; ; \;\; R \leq \frac{\pi}{2\sqrt{\beta}}$$

◇ Hadamard-von Mangoldt : squared distance is globally strongly convex on simply-connected surfaces of negative curvature

Convexity of squared distance



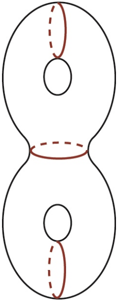
◇ Recall the Hessian :

$$\nabla^2 e_x \simeq \begin{pmatrix} 2 & 0 \\ 0 & 2rS \end{pmatrix} \geq 2\,\beta r \cot(\beta r) \qquad \text{(positive curvature)}$$

$$\geq 2 \qquad \text{(non-positive curvature)}$$

◇ But is it convex ?

The problem comes from closed geodesics



◇ Convexity radius : when is a geodesic ball convex ?

$i(M) = \inf_x i(x)$
(possibly zero)

$B(x, R)$ is convex if $R \leq \dfrac{i(M)}{2}$; $R \leq \dfrac{\pi}{2\sqrt{\beta}}$

◇ Hadamard-von Mangoldt : squared distance is globally strongly convex on simply-connected surfaces of negative curvature

Plan

Riemannian connection (Levi-Civita)

◇ Why do we need a connection ?

Define acceleration of a curve, Hessian of a function, ... : differentiate a vector field !

◇ Affine connection :

$$\nabla_v X \;=\; \text{derivative of vector field } X \text{ along direction } v$$

◇ Levi-Civita connection : (compatible with the metric)

$$\nabla_v \langle X, Y\rangle \;=\; \langle \nabla_v X, Y\rangle + \langle X, \nabla_v Y\rangle$$

(zero torsion)

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

◇ Koszul formula :

$$\nabla Y = \underbrace{\frac{1}{2}\mathcal{L}_Y g}_{\text{elasticity tensor}} + \underbrace{\frac{1}{2} d(g \cdot Y)}_{\text{skew-symmetric}}$$

Riemannian connection (Levi-Civita)

◇ Why do we need a connection ?

Define acceleration of a curve, Hessian of a function, ... : differentiate a vector field !

◇ Affine connection :

$$\nabla_v X \;=\; \text{derivative of vector field } X \text{ along direction } v$$

◇ Levi-Civita connection : (compatible with the metric)

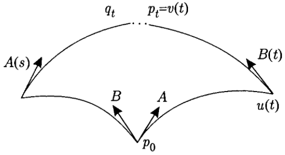
$$\nabla_v \langle X, Y\rangle \;=\; \langle \nabla_v X, Y\rangle + \langle X, \nabla_v Y\rangle$$

(zero torsion)

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

◇ Koszul formula :

$$\nabla Y = \underbrace{\frac{1}{2}\mathcal{L}_Y g}_{\text{elasticity tensor}} + \underbrace{\frac{1}{2}d(g\cdot Y)}_{\text{skew-symmetric}}$$



Riemannian connection (Levi-Civita)

◇ Example 1 : embedded submanifold

$$(\nabla_v X)(x) = \Pi_x(D_v X) \quad \text{projection of component-wise derivative}$$

◇ Example 2 : geodesic spherical coordinates

$$\nabla_{\partial_r}\partial_r = 0 \qquad \nabla_{\partial_r}\partial_\theta = S(r,\theta)\partial_\theta$$

$$\nabla_{\partial_\theta}\partial_r = S(r,\theta)\partial_\theta \qquad \nabla_{\partial_\theta}\partial_\theta = -(f^2)_r\partial_r$$

◇ Interpretation :

$$\underbrace{\nabla_{\partial_r}\partial_r = 0}_{\text{geodesic equation}} \qquad \underbrace{\nabla_{\partial_\theta}\partial_r = S(r,\theta)\partial_\theta}_{\text{Hessian of distance}}$$

Riemannian connection (Levi-Civita)

⋄ Example 1 : embedded submanifold

$$(\nabla_v X)(x) = \Pi_x(D_v X) \quad \text{projection of component-wise derivative}$$

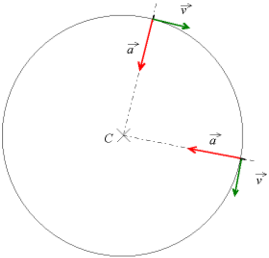
⋄ Example 2 : geodesic spherical coordinates

$$\nabla_{\partial_r}\partial_r = 0 \qquad \nabla_{\partial_r}\partial_\theta = S(r,\theta)\partial_\theta$$

$$\nabla_{\partial_\theta}\partial_r = S(r,\theta)\partial_\theta \qquad \nabla_{\partial_\theta}\partial_\theta = -(f^2)_r\partial_r$$

⋄ Interpretation :

$$\underbrace{\nabla_{\partial_r}\partial_r = 0}_{\text{geodesic equation}} \qquad \underbrace{\nabla_{\partial_\theta}\partial_r = S(r,\theta)\partial_\theta}_{\text{Hessian of distance}}$$



Some definitions

⋄ Geodesic equation :

$$\nabla_{\dot\gamma}\dot\gamma = 0 \qquad \text{zero-acceleration parameterised curve}$$

⋄ Gradient ∇f :

$$\underbrace{v \cdot f}_{\text{directional derivative}} = \underbrace{df(x) \cdot v}_{df(x): T_x M \to M} = \underbrace{\langle \nabla f, v \rangle_x}_{\text{gradient}}$$

⋄ Hessian $\nabla^2 f$:

$$\underbrace{\nabla^2 f \cdot v = \nabla_v \nabla f(x)}_{\text{self-adjoint endomorphism}} \qquad \underbrace{\nabla^2 f \cdot (u, v) = \langle \nabla^2 f \cdot u, v \rangle}_{\text{symmetric bilinear form}}$$

⋄ Second-order Taylor :

$$f(\gamma(1)) = f(\gamma(0)) + \langle \nabla f, \dot\gamma \rangle_{\gamma(0)} + \frac{1}{2} \nabla^2 f \cdot (\dot\gamma, \dot\gamma) \Big|_{t^*} \qquad \text{where } t^* \in (0, 1)$$

The exponential map

◇ Definition from ODE theory :

$$\underbrace{\nabla_{\dot\gamma}\dot\gamma = 0 \ \text{ i.c. } \left(\gamma(0) = x \ \text{ and } \ \dot\gamma(0) = v = y^i\partial_i\right)}_{\text{the solution is unique for given initial conditions}} \implies \mathrm{Exp}_x(v) = \gamma(1) = y$$

◇ Completeness and Hopf-Rinow :

$\mathrm{Exp}_x(v)$ is defined for all $v \iff$ any $x, y \in M$ connected by a minimising geodesic

◇ Normal coordinates : is the relation $y \mapsto y^i$ unique ?

$M = D_x \cup \mathrm{Cut}(x)$ Exp diffeomorphism of D_x

◇ What happens on $\mathrm{Cut}(x)$:

— $|d\mathrm{Exp}_x(v)| = 0$ conjugate point

— Exp is not bijective

The exponential map

◇ Definition from ODE theory :

$$\underbrace{\nabla_{\dot\gamma}\dot\gamma = 0 \ \text{ i.c. } \left(\gamma(0) = x \ \text{ and } \ \dot\gamma(0) = v = y^i\partial_i\right)}_{\text{the solution is unique for given initial conditions}} \implies \mathrm{Exp}_x(v) = \gamma(1) = y$$

◇ Completeness and Hopf-Rinow :

$\mathrm{Exp}_x(v)$ is defined for all $v \iff$ any $x, y \in M$ connected by a minimising geodesic

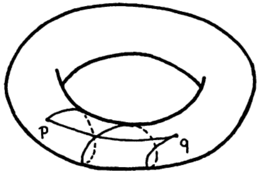
◇ Normal coordinates : is the relation $y \mapsto y^i$ unique ?

$M = D_x \cup \mathrm{Cut}(x)$ Exp diffeomorphism of D_x

◇ What happens on $\mathrm{Cut}(x)$:

— $|d\mathrm{Exp}_x(v)| = 0$ conjugate point

— Exp is not bijective



The exponential map

◇ Definition from ODE theory :

$$\underbrace{\nabla_{\dot\gamma}\dot\gamma = 0 \ \text{ i.c. } \left(\gamma(0) = x \ \text{ and } \ \dot\gamma(0) = v = y^i\partial_i\right)}_{\text{the solution is unique for given initial conditions}} \implies \mathrm{Exp}_x(v) = \gamma(1) = y$$

◇ Completeness and Hopf-Rinow :

$$\mathrm{Exp}_x(v) \text{ is defined for all } v \iff \text{any } x, y \in M \text{ connected by a minimising geodesic}$$

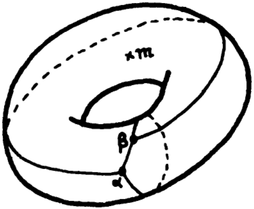
◇ Normal coordinates : is the relation $y \mapsto y^i$ unique ?

$$M = D_x \cup \mathrm{Cut}(x) \qquad\qquad \mathrm{Exp} \text{ diffeomorphism of } D_x$$

◇ What happens on $\mathrm{Cut}(x)$:

— $|d\mathrm{Exp}_x(v)| = 0$ conjugate point

— Exp is not bijective



Convex sets and functions

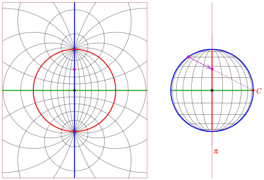
◇ Convex $A \subset M$:

$$x, y \in A : \text{ length-minimising } \gamma(0) = x, \gamma(1) = y \text{ and } \gamma(t) \in A$$

A ball may fail to be convex!!

◇ Convexity radius : (small balls are always convex)

$$R_{cx}(M) \geq \min\left\{\frac{i(M)}{2}, \frac{\pi}{2\sqrt{\beta}}\right\}$$



◇ Convex function $f : A \to R$:

$$f \circ \gamma : [0, 1] \to \mathbb{R} \text{ is convex}$$

◇ Example :

$$e_x : B(x, R) \to \mathbb{R} \text{ where } e_x(y) = d^2(x, y)$$

$\rightsquigarrow$ if $R < R_{cx}$ this function is convex

◇ Characterisation :

$$A \text{ is convex and } \nabla^2 \varphi(y) \geq 0 \text{ for } y \in A$$

Hadamard manifolds

M simply connected, complete, with sectional curvature ≤ 0

- Examples :
  - Euclidean space
  - Poincaré half plane
  - Cones of covariance matrices
- Nice properties : $i(M) = \infty \rightsquigarrow$ no conjugate points, no closed geodesics
  - Exp is a diffeomorphism
  - squared distance is smooth
  - squared distance is strongly convex

$$\nabla^2 e_x(y) \geq 1 \quad \text{all } x \text{ and } y \text{ in } M \qquad\qquad (e_x(y) = d^2(x, y))$$

- Barycentre problem : (Fréchet mean)

$$\underbrace{F(y) = \frac{1}{N}\sum_{n=1}^{N} d^2(x_n, y)}_{\text{smooth strongly convex function}} \rightsquigarrow \text{unique minimum and stationary point } \hat{x}_N$$

Plan

Stochastic optimisation

◇ Loss function ($L : M \longrightarrow \mathbb{R}$) :

$$L(x) = \mathbb{E}_z \, \ell(x, z) \quad \text{or} \quad L(x) = \frac{1}{N} \sum_{n=1}^{N} \ell(x, z_n)$$

◇ Main issues :

loss function unknown ; evaluation too costly

◇ Idea : learn and optimise at the same time!!

– generate or observe z_n where $n = 1, 2, \ldots$

– follow the gradient on average $\quad x_{n+1} = \mathrm{Exp}_{x_n}\left(-\gamma_{n+1} \nabla \ell(x_n, z_{n+1})\right)$

◇ Deterministic vs stochastic :

	find stationary point	find local min	local rate of convergence
deterministic	YES	NO	geometric
stochastic	YES	YES	harmonic

◇ Limit set : connected component of $\{\nabla L = 0\} \cup \{\infty\}$ (for more, recall the capture theorem)

Local rate of convergence

$$x_{n+1} = \mathrm{Exp}_{x_n}(-\gamma_{n+1} \nabla \ell(x_n, z_{n+1}))$$

⋄ Assumptions :

- $(x_n) \subset D$ compact convex set
- exactly one stationary point $x^* \in D$
- L is μ-strongly convex in D
- controle of moments of noise

strong convexity $(x, y \in D)$: $\underbrace{L(y) - L(x) \geq \langle \nabla L(x), \mathrm{Exp}_x^{-1}(y) \rangle_x + \frac{\mu}{2} d^2(x, y)}_{\text{(a convex function is above all its tangents)}}$

$\rightsquigarrow$ strong attraction : $\underbrace{-\frac{\mu}{2} d^2(x, x^*) \geq \langle \nabla L(x), \mathrm{Exp}_x^{-1}(x^*) \rangle_x}_{\text{(attraction to } x^* \text{ is super-linear)}}$

Local rate of convergence

$$x_{n+1} \; = \; \mathrm{Exp}_{x_n}\left(-\gamma_{n+1}\nabla\ell(x_n, z_{n+1})\right) \rightsquigarrow \textit{set } u_{n+1} = \gamma_{n+1}\nabla\ell(x_n, z_{n+1})$$

◇ Assumptions :

- $(x_n) \subset D$ compact convex set
- exactly one stationary point $x^* \in D$
- L is μ-strongly convex in D
- controle of moments of noise

triangle comparison : $d^2(x_{n+1}, x^*) \leq d^2(x_n, x^*) + 2\langle u_{n+1}, \mathrm{Exp}^{-1}_{x_n}(x^*)\rangle + DS_\alpha(D)\|u_{n+1}\|^2$

conditional expectation : $\mathbb{E}_n\, d^2(x_{n+1}, x^*) \leq d^2(x_n, x^*) + 2\gamma_{n+1}\langle \nabla L(x_n), \mathrm{Exp}^{-1}_{x_n}(x^*)\rangle + C\gamma^2_{n+1}$

strong attraction : $\mathbb{E}_n\, d^2(x_{n+1}, x^*) \leq (1-\gamma_{n+1}\mu)\, d^2(x_n, x^*) + C\gamma^2_{n+1}$

take expectation : $\mathbb{E}\, d^2(x_{n+1}, x^*) \leq (1-\gamma_{n+1}\mu)\,\mathbb{E}\, d^2(x_n, x^*) + C\gamma^2_{n+1}$

◇ A first conclusion :

we must take $\limsup \gamma_{n+1}\mu < 1$

The problem of tuning

◇ Usual choice of step-size :

$$\gamma_n = \frac{A}{n^\alpha + B} \quad \text{where} \quad \alpha \in (0,1] \qquad \alpha \uparrow 1 \text{ stops the algorithm faster}$$

◇ Local rate of convergence :

$$\mathbb{E}\, d^2(x_n, x^*) \le C\gamma_n^\beta \quad \text{where} \quad \beta \in (0,1)$$

◇ Optimal rate :

$$\mathbb{E}\, d^2(x_n, x^*) \le C\gamma_n \quad \text{requires} \quad A > \frac{\alpha}{\mu}$$

(Please note these are only local rates!)

◇ Conclusion : ∗ we need to know μ (spend money)

∗ we need to guess μ (spend time)

∗ convergence can be arbitrarily bad

∗ anyway, a small μ is a bad case

◇ Can we get around knowing μ ?

$\rightsquigarrow$ there exist some very nice tricks

Averaged stochastic gradient

⋄ Maintain a constant step-size :

$$x_{n+1} = \mathrm{Exp}_{x_n}(-\gamma\,\nabla\ell(x_n, z_{n+1})) \qquad \gamma \text{ constant (or slowly decreasing)}$$

⋄ Does this converge ?

a stationary Markov process (the question is convergence in law, or ergodicity)

⇝ somehow, we need to stabilise it

⋄ Recursive Riemannian average : (generalise the Polyak average)

$$\underbrace{\hat{x}_{n+1} = \hat{x}_n \,\#_{\frac{1}{n+1}}\, x_{n+1}}_{\text{geodesic weighted average}}$$

In a Euclidean space, this reduces to

$$\hat{x}_{n+1} = \frac{n}{n+1}\hat{x}_n + \frac{1}{n+1}x_{n+1}$$

⇝ SGD becomes the input of a Riemannian AR(1)

A digression about barycentres

⋄ Riemannian barycentre (Fréchet mean) :

$$\bar{x} \text{ any global minimum of } \mathcal{E}(x) = \int_M d^2(x,y)P(dy)$$

apparently, just stochastic optimisation

⋄ In Euclidean space :

$$\bar{x} = \int_M y\,P(dy) \qquad \text{unique global minimiser}$$

⋄ Law of large numbers :

$$\hat{x}_n = \frac{1}{n}\sum_{m=1}^{n} x_m \to \bar{x} \qquad \hat{x}_{n+1} = \frac{n}{n+1}\hat{x}_n + \frac{1}{n+1}x_{n+1}$$

⋄ General Riemannian case :

$\mathcal{E}(x)$ is non-differentiable, non-convex, and has multiple minima !!

– Conclsuion :

Open problem : barycentre of a Markov chain

Plan

The problem of recursive estimation

◇ Estimation/learning problem :

minimise a statistical divergence $\theta^* = \operatorname{argmin} \mathrm{D}(\mathrm{P}_{\text{true}}|\mathrm{P}_\theta)$

$\theta \in \Theta$ (model space, a manifold)

no reason to think $\mathrm{P}_{\text{true}} = \mathrm{P}_{\theta^*}$

$\rightsquigarrow$ but we do not know P_{true} in the first place!!

◇ Empirical estimation : (example of KL divergence)

$$\mathrm{D}(\mathrm{P}_{\text{true}}|\mathrm{P}_\theta) = \int \log\left[\frac{\mathrm{p}_{\text{true}}(x)}{\mathrm{p}_\theta(x)}\right] d\mathrm{P}_{\text{true}}(x) \approx \underbrace{\frac{1}{N}\sum_{n=1}^{N}\log \mathrm{p}_{\text{true}}(x_n) - \frac{1}{N}\sum_{n=1}^{N}\log \mathrm{p}_\theta(x_n)}_{\text{first term does not depend on } \theta}$$

Drawbacks	Advantages
changes the original minimisation problem	consistent, asymptotically efficient
recomputes from scratch with new samples	uses established optimisation methods
not suitable to very complicated models	

◇ Recursive estimation : we try to have the same advantages without the drawbacks

The problem of recursive estimation

◇ Estimation/learning problem :

minimise a statistical divergence $\theta^* = \text{argmin}\, \mathrm{D}(\mathrm{P}_{\text{true}} | \mathrm{P}_\theta)$

$\theta \in \Theta$ (model space, a manifold)

no reason to think $\mathrm{P}_{\text{true}} = \mathrm{P}_{\theta^*}$

⇝ but we do not know P_{true} in the first place!!

◇ Empirical estimation : (example of KL divergence)

$$\mathrm{D}(\mathrm{P}_{\text{true}} | \mathrm{P}_\theta) = \int \log \left[\frac{\mathrm{p}_{\text{true}}(x)}{\mathrm{p}_\theta(x)} \right] d\mathrm{P}_{\text{true}}(x) \approx \underbrace{\frac{1}{N} \sum_{n=1}^{N} \log \mathrm{p}_{\text{true}}(x_n) - \frac{1}{N} \sum_{n=1}^{N} \log \mathrm{p}_\theta(x_n)}_{\text{first term does not depend on } \theta}$$

◇ Recursive estimation : we try to have the same advantages without the drawbacks

$$\theta_{n+1} = \varphi(\theta_n, \ldots, x_{n+1}) \qquad \lim \theta_n = \theta^*$$

The Fisher information metric

◇ First definition :

a metric adapted to the divergence $\quad \mathrm{D}(\mathrm{P}_\theta|\mathrm{P}_{\theta+d\theta}) = \frac{1}{2}\,\|d\theta\|^2 + \dots$

◇ Is this really a metric ?

$$\text{(case of Kullback-Leibler)}\ \ \|d\theta\|^2 = -\sum_a \sum_b \mathbb{E}_\theta\left(\frac{\partial^2 \log \mathrm{p}_\theta}{\partial\theta^a \partial\theta^b}\right) d\theta^a d\theta^b$$

$$\text{Rao's discovery} \qquad \underbrace{\|d\theta\|^2 = \|d\theta'\|^2}_{\text{invariance by reparameterisation}}$$

◇ Second definition (Chentsov's theorem) : a formula is not a definition

there is (essentially) a unique metric on Θ invariant by sufficient statistics

$$\mathrm{D}(\mathrm{P}_\theta|\mathrm{P}_{\theta+d\theta}) = \mathrm{D}(\mathrm{P}_\theta \circ \varphi|\mathrm{P}_{\theta+d\theta} \circ \varphi) \qquad (\varphi \text{ sufficient statistic})$$

⇝ many computations become automatic . . .

⇝ explains the appearance of affine-invariance

Recursive estimation

◇ Gradient flow :

$$\dot{\theta} \;=\; -\nabla_\theta\, \mathrm{D}(\mathrm{P}_{\mathrm{true}}|\mathrm{P}_\theta)$$

Limit set (forward)

$\{\infty\}$ or $\{$stationary point$\}$ or $\{$stationary infinite set$\}$

$\rightsquigarrow$ we cannot run this dynamical system !!

◇ Stochastic approximation :

$$\theta_{n+1} \;=\; \mathrm{Exp}_{\theta_n}\left(\gamma_{n+1}\, u(\theta_n, x_{n+1})\right)$$

◇ Limit set (a.s.) :

$$\left.\begin{array}{l} \sum \gamma_n = \infty\,,\;\; \sum \gamma_n^2 < \infty \\ \mathbb{E}_{\mathrm{true}}\, u(\theta, x) = -\nabla \mathrm{D}(\theta) \end{array}\right\} \Longrightarrow \text{same as above}$$

◇ Reflected algorithm : introduce “walls” to avoid going to $\{\infty\}$

◇ Unstable points :

$$\varepsilon_n(\theta) \;=\; u(\theta, x) - \mathbb{E}_{\mathrm{true}}\, u(\theta, x) \text{ (approximation noise)}$$

isotropic noise $\Longrightarrow$ $\mathbb{P}_{\mathrm{true}}(\theta_n \to \text{unstable point}) \;=\; 0$

Recursive estimation

◇ Gradient flow :

$$\dot{\theta} = -\nabla_\theta \, \mathrm{D}(\mathrm{P}_{\mathrm{true}} | \mathrm{P}_\theta)$$

Limit set (forward)

$\{\infty\}$ or {stationary point} or {stationary infinite set}

$\rightsquigarrow$ we cannot run this dynamical system !!

◇ Stochastic approximation :

$$\theta_{n+1} = \mathrm{Exp}_{\theta_n}(\gamma_{n+1} \, u(\theta_n, x_{n+1}))$$ *or any C^2 retraction*

◇ Limit set (a.s.) :

$$\left.\begin{array}{l} \sum \gamma_n = \infty \,, \ \sum \gamma_n^2 < \infty \\ \mathbb{E}_{\mathrm{true}} \, u(\theta, x) = -\nabla \mathrm{D}(\theta) \end{array}\right\} \Longrightarrow \text{same as above}$$

◇ Reflected algorithm : introduce “walls” to avoid going to $\{\infty\}$

◇ Unstable points :

$$\varepsilon_n(\theta) = u(\theta, x) - \mathbb{E}_{\mathrm{true}} \, u(\theta, x) \text{ (approximation noise)}$$

isotropic noise $\Longrightarrow \mathbb{P}_{\mathrm{true}}(\theta_n \longrightarrow \text{unstable point}) = 0$

Local rate of convergence

◇ Behavior at stable point :

$$\nabla D(\theta^*) = 0 \quad ; \quad \underbrace{\nabla^2 D(\theta^*) \succ 0 \text{ (least eigenvalue } \lambda)}_{\lambda \text{ depends on the choice of metric}}$$

◇ Strong attraction : if $\lambda > \mu > 0$ there exists open Θ^* at θ^*

$$-\mu\, d^2(\theta,\theta^*) \geq \langle \nabla D(\theta), \mathrm{Exp}_\theta^{-1}(\theta^*)\rangle_\theta \qquad \text{for all } \theta \in \Theta^*$$

◇ Best achievable rate :

$$\gamma_n = \frac{A}{n} \text{ and } A > \frac{1}{2\mu} \implies d^2(\theta,\theta^*) = O\left(n^{-1}\right)$$

◇ Automatic tuning : (assume $\theta^* = \theta_{\text{true}}$)

$$\text{information metric} \rightsquigarrow \lambda = 1$$

Local rate of convergence

⋄ Behavior at stable point :

$$\nabla D(\theta^*) = 0 \quad ; \quad \underbrace{\nabla^2 D(\theta^*) \succ 0 \text{ (least eigenvalue } \lambda)}_{\lambda \text{ depends on the choice of metric}} \quad \textit{(Hessian as bilinear form)}$$

⋄ Strong attraction : if $\lambda > \mu > 0$ there exists open Θ^* at θ^*

$$-\mu\, d^2(\theta, \theta^*) \geq \langle \nabla D(\theta), \mathrm{Exp}_\theta^{-1}(\theta^*) \rangle_\theta \qquad \text{for all } \theta \in \Theta^*$$

⋄ Best achievable rate :

$$\gamma_n = \frac{A}{n} \text{ and } A > \frac{1}{2\mu} \implies d^2(\theta, \theta^*) = O\left(n^{-1}\right)$$

⋄ Automatic tuning : (assume $\theta^* = \theta_{\text{true}}$)

$$\text{information metric} \rightsquigarrow \lambda = 1$$

Asymptotic normality

◇ Normalised error :

$$\xi_n = \sqrt{\gamma_n}\, \mathrm{Exp}^{-1}_{\theta^*}(\theta_n)$$

◇ The CLT :

$$\xi_n \implies \mathcal{N}(0\,,\,\Sigma) \text{ (expressed in o.n.b.)}$$

◇ Lyapunov equation :

$$H\,\Sigma + \Sigma H = -A^2\,\Sigma^* \qquad \begin{aligned} \gamma_n &= \tfrac{A}{n} \\ H &= \tfrac{1}{2}\,\mathrm{Id} - \nabla^2 D(\theta^*) \\ \Sigma^* &= \mathbb{E}_{\text{true}}\left(\varepsilon_n(\theta^*) \otimes \varepsilon_n(\theta^*)\right) \end{aligned}$$

◇ What does this mean ??!

$$\text{(asymptotic behavior)} \qquad \underbrace{d\xi(t) = H\xi(t)dt + \Sigma^{-\frac{1}{2}}\, dW(t)}_{\text{linear attraction + white noise}}$$

◇ Asymptotic efficiency

$$\text{information metric} \quad \rightsquigarrow \quad \Sigma = \Sigma^* = (I(\theta^*))^{-1}$$

$$\underbrace{d(\theta_n\,,\,\theta^*) \Rightarrow \chi^2_{\dim\Theta}}_{\text{useful for change detection}}$$

Unknown information metric

⋄ Examples of "difficult models" :

- mixture models
- neural networks
- FIM known but complicated

⋄ "partial FIM" :

retain diagonal part and try to find λ and Σ

⋄ Automatic tuning : (averaged stochastic gradient under some suitable metric)

$$\theta_{n+1} = \mathrm{Exp}_{\theta_n}(-\gamma\, u(\theta_n, x_{n+1})) \qquad \gamma \text{ constant (or slowly decreasing)}$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n \,\#_{\frac{1}{n+1}}\, \hat{\theta}_{n+1} \qquad \text{geodesic average}$$

⋄ this guarantees $O\left(n^{-1}\right)$ convergence rate and asymptotic efficiency

⋄ Exp and # need to be manageable (chose a symmetric geometry ..)

Plan

1. Introduction with surfaces
2. Higher dimension
3. Convex stochastic optimisation
4. Riemannian recursive estimation
5. Proposed reading

Proposed reading

◇ Manifolds and Riemannian geometry :

J.M. Lee : Introduction to Topological manifolds

J.M. Lee : Introduction to Smooth Manifolds

J.M. Lee : Introduction to Riemannian manifolds

◇ Information geometry :

S.I Amari : Methods of information geometry

... ... : Learn from the state of the art !!

◇ Recursive estimation :

Nevilson & Hasminskii : Stochastic approximation and recursive estimation

Marie Duflo : Algorithmes Stochastiques + Random iterative models

◇ Riemannian recursive estimation :

Bonnabel : Stochastic gradient descent on Riemannian manifolds

Tripuraneni & *al* : Averaging stochastic gradient descent on Riemannian manifolds