Information Geometry

B.Khodabandeh, S.Heidari, D.Zinati

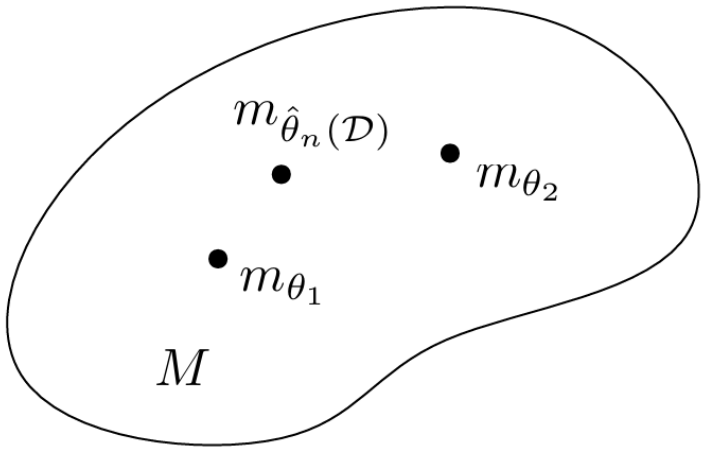
Sharif University of Technology
Electrical Engineering Department

Spring 2024

Overview

Introduction

Geometric Structure of Probability Distributions

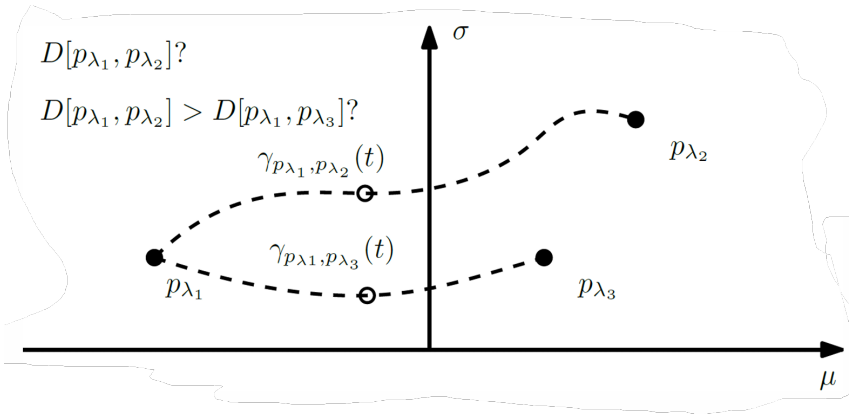


Studying the properties of a parameterized set can naturally lead to manifolds' geometry since it is closely related to the definition of manifolds which consists of homeomorphisms between open subsets to $\mathbb{R}^n$(the space of parameters).

Consider the set of normal distributions on scalars:

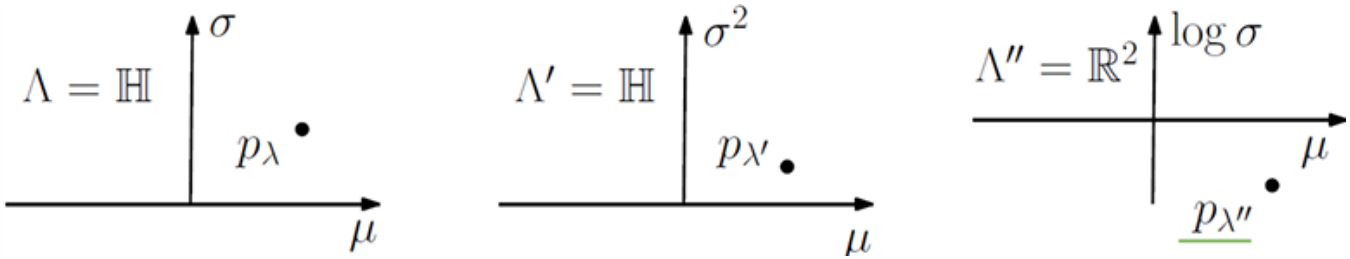
$$\mathcal{P} = \left\{ p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right),\ \theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{++} \right\}$$

Here θ is the parameter of the distribution which is inside the upper half-plane. We can consider this set as a manifold.



Some questions arising from considering this manifold are: How to interpolate between two normal distributions? How to define a distance between them?

By viewing the set $\mathcal{P}$ as a manifold we can think of the points of the set(distributions) independent of the parameterization. This approach leads to some invariance principles which have important meanings and can be useful in some ways.

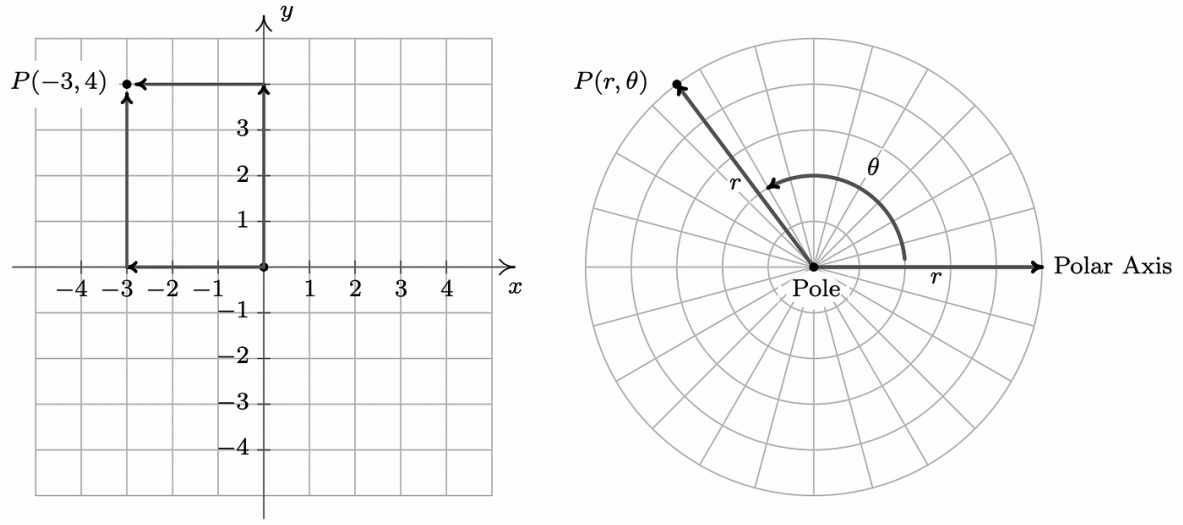


$$\mathcal{P} = \left\{ p_{\lambda''}(x) = \frac{1}{\sqrt{2\pi e^{\lambda_2''}}} \exp\left(-\frac{(x-\lambda_1'')^2}{2e^{2\lambda_2''}} \right), \lambda = (\mu, \log(\sigma)) \in \mathbb{R}^2 \right\}$$

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

Here $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$ is the upper half-plane.

For example, the 2-D Euclidean space:



here, both coordinates represent the same phenomenon, but they are expressed differently.
It is intuitive that our notion of distance should not depend on our parameterization, and only on the underlying model.
and if we need to interpolate between two points, this paths should also not depend on the coordinates.

Notation

Notation	Description
$\mathcal{M}, \mathcal{N}, \mathcal{P}$	Manifold
$T_P\mathcal{M}$	Tangent space of manifold $\mathcal{M}$ at point P
$T_P^*\mathcal{M}$	Cotangent space of manifold $\mathcal{M}$ at point P
X, Y, Z	Vector or Vector Field
ω, α, β	Covector or Covector Field
$\otimes$	Tensor Product
$\{e_i\}_{i=1}^n$	Basis
$\{e^i\}_{i=1}^n$	Covector Basis
$\{\partial_i\}_{i=1}^n$	Coordinate Basis
$\{d\theta^i\}_{i=1}^n$	Coordinate Covector Basis
g	Metric Tensor
g_{ij}	(i,j)th Element of the Metric Tensor
∇	Affine Connection

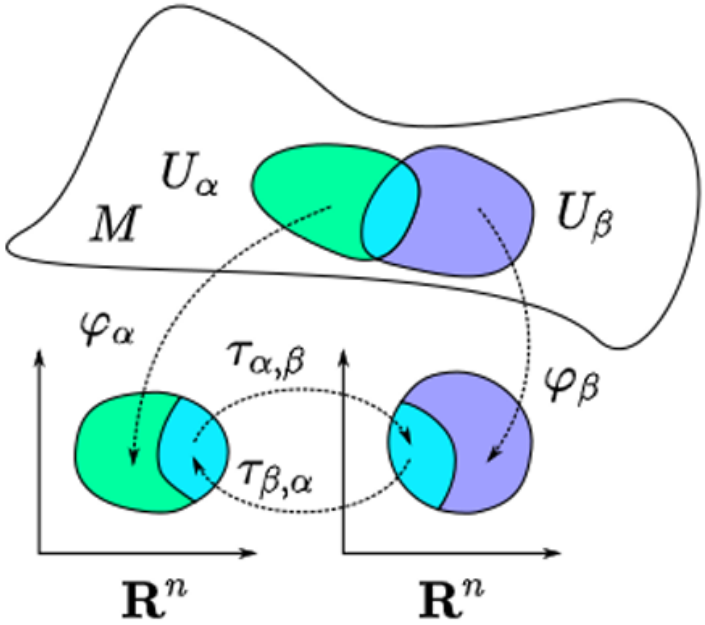
Notation

Notation	Description
$\nabla_{(.)}(.)$	Covariant Derivative
${}^{LC}\nabla$	Levi-Civita Connection
Γ^{i}_{jk}	Connection Coefficients
$\prod_{C}^{\nabla}$	Parallel Transport
R	Riemann Tensor
R^{i}_{jkl}	Riemann Tensor Components
C	Amari-Chentsov totally symmetric tensor
$\mathbb{E}$	Mathematical Expectation
$I(\theta)$	Fisher Information Matrix
${}^{*}\nabla$	Conjugate Connection
${}^{\alpha}\nabla$, ${}^{-\alpha}\nabla$	alpha Connection
${}^{e}\nabla$, ${}^{m}\nabla$	e Connection, m Connection

Preliminaries

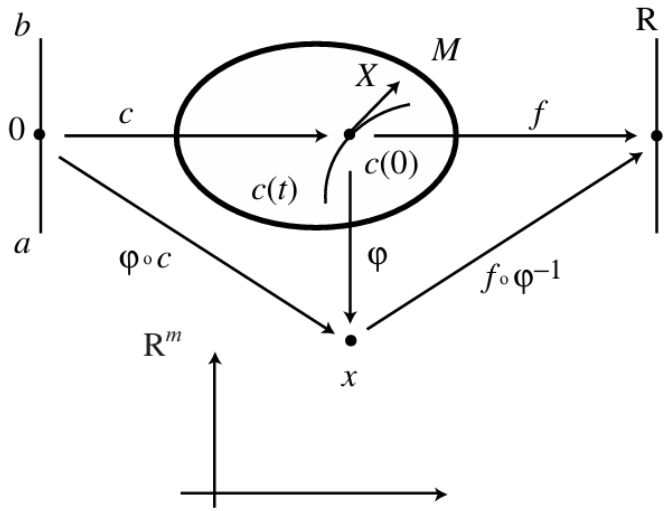
Differential Geometry Background

What is a Manifold?



A Manifold is a Topological Space(equipped with a Hausdorff topology) that is covered by open sets homeomorphic to subsets of $\mathbb{R}^n$ called charts. In the intersections, the charts should be diffeomorphic.

How to define vectors and covectors?



There are two equivalent ways to define Tangent Vector Spaces on Manifolds:

- By using the equivalence Classes of curves that pass through a point
- Defining vectors as first-order differential operators on functions.

The second definition is more suitable for our purposes. The Cotangent Space is just the dual space of the Tangent Space

Riemannian Geometry Background

The Metric Tensor:

$$g = g_{ij} d\theta^i \otimes d\theta^j \in T_P^* \mathcal{M} \otimes T_P^* \mathcal{M}$$

The Metric Tensor can be represented with a positive definite matrix. This Metric Tensor is the definition of the inner product between vectors.

$$\langle X, Y \rangle := g(X, Y)$$

It is also used for defining the infinitesimal length between points on the Manifold:

$$ds^2 = g_{ij} d\theta^i \ d\theta^j$$

Affine Connection(∇): It is defined to let us make connections between Tangent Spaces. It allows us to define concepts such as Covariant Derivative, Geodesics, and Curvature.

The Covariant Derivative of Y with respect to X:

$$\nabla_X Y = X^j(\partial_j Y^i + \Gamma^i{}_{jk} Y^k)\partial_i$$

Levi-Civita Connection($^{LC}\nabla$): Is the connection induced by the metric tensor which preserves the Metric tensor in Covariant Derivatives
Geodesics(Auto-Parallel Curves):

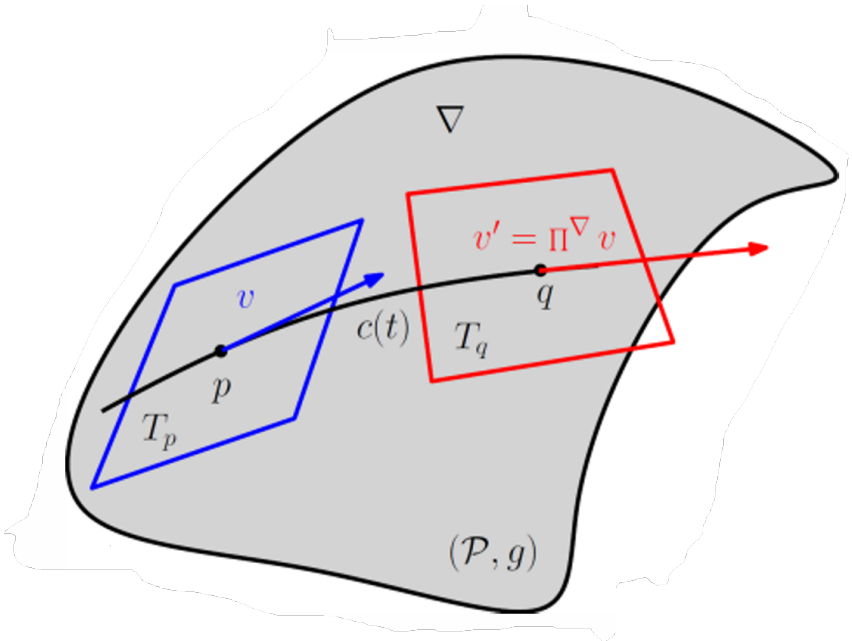
$$\nabla_{\dot\gamma(t)}\dot\gamma(t) \propto \dot\gamma(t)$$

Which can be written as below:

$$\ddot\theta^i + \Gamma^i{}_{jk}\dot\theta^j\dot\theta^k \propto \dot\theta^i$$

Geodesics

Geodesics are curves that generalize the concept of straight lines to curved spaces. They can be visualized as below:

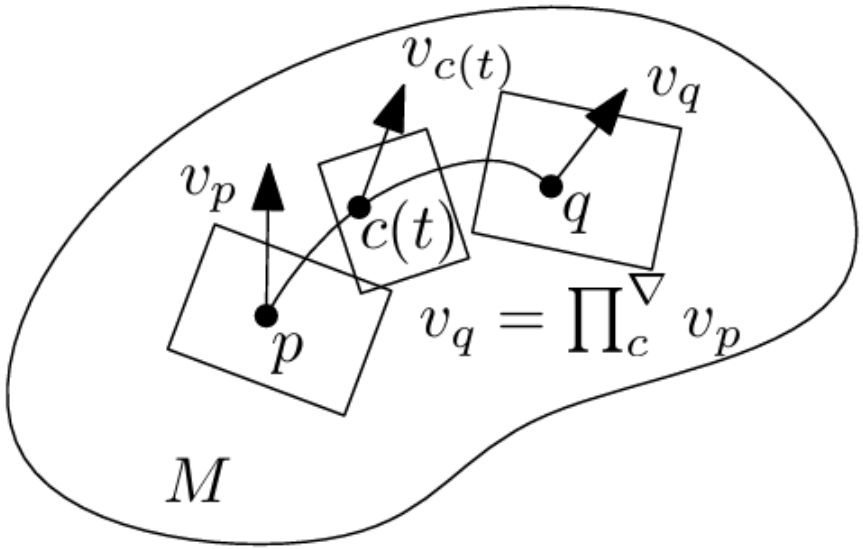


Parallel Transport

- Parallel transport allows the comparison of vectors at different points on a manifold.
- A vector $v \in T_pM$ is parallel transported along a curve $c(t)$ if:

$$\frac{D}{dt}v(t) = 0, \tag{1}$$

where $\frac{D}{dt}$ is the covariant derivative along the curve $c(t)$.



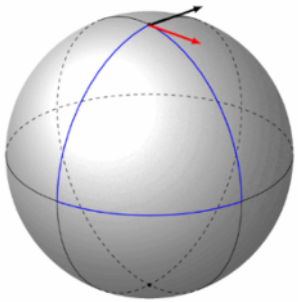
Riemannian Curvature

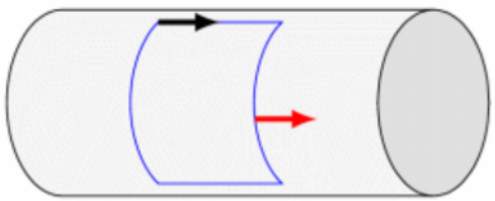
- The curvature of a manifold is captured by the Riemann curvature tensor R.
- For vector fields X, Y, Z:

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

- In local coordinates, the components R^l_{ijk} of the Riemann tensor are given by:

$$R^l_{ijk} = \partial_j \Gamma^l_{ik} - \partial_k \Gamma^l_{ij} + \Gamma^l_{jm}\Gamma^m_{ik} - \Gamma^l_{km}\Gamma^m_{ij}.$$





Dual Structure Manifolds

Conjugate Connection Manifolds

- Definition: A connection ∇^* is conjugate to a connection ∇ with respect to a metric tensor g if:

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle \tag{2}$$

for any vector fields X, Y, Z.

- Conjugate connections (∇, ∇^*) preserve the metric under dual parallel transport.
- The mean connection $\overline{\nabla}$:

$$\overline{\nabla} = \frac{1}{2}(\nabla + \nabla^*) \tag{3}$$

coincides with the Levi-Civita connection.

Statistical Manifolds

- Definition: A statistical manifold (M, g, C) consists of a manifold M with a metric tensor g and a symmetric cubic tensor C.
- The cubic tensor C is defined as:

$$C_{ijk} = \Gamma_{ijk} - \Gamma^*_{ijk} \tag{4}$$

where Γ_{ijk} and Γ^*_{ijk} are the Christoffel symbols for ∇ and ∇^* respectively.

- Statistical manifolds are essential in defining a family of conjugate connections.

Family of Conjugate Connection Manifolds (α)

- For any pair of conjugate connections (∇, ∇^*), a 1-parameter family ∇^α can be defined:

$$\nabla^\alpha = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^* \tag{5}$$

- The α-connection ∇^α can also be expressed as:

$$\nabla^\alpha_X Y = \nabla_X Y + \frac{\alpha}{2}C(X,Y) \tag{6}$$

where C is the cubic tensor.

The Fundamental Theorem of Information Geometry

- Theorem: If a torsion-free affine connection ∇ has constant curvature κ, then its conjugate connection ∇^* also has constant curvature κ.
- This property leads to dually flat manifolds, where both ∇ and ∇^* are flat.
- Example: In a dually flat manifold, the geodesics corresponding to ∇ and ∇^* are both straight lines in their respective coordinate systems.

Dually Flat Manifolds

- Definition: A manifold (M, g, ∇, ∇^*) is dually flat if both ∇ and ∇^* are flat.
- Dually flat manifolds allow for a global affine coordinate system.
- Example: Exponential family of probability distributions often forms a dually flat manifold.

Divergences

Divergences

Loosely speaking, a divergence $D : M \times M \to [0, \infty)$ is a smooth distance, potentially asymmetric.

- $D(\theta : \theta') \geq 0$, equality $\iff \theta - \theta'$
- $\partial_{i,.} D(\theta : \theta') = \partial_{.,j} D(\theta : \theta') = 0$
- $-\partial_{i,.} \partial_{.,j} D(\theta : \theta')$ is positive definite

$$\Downarrow$$

Acts like $distance^2$

The dual divergence is defined by swapping the arguments:

$$D^*(\theta : \theta') = D(\theta' : \theta)$$

Conjugate Connections from Divergences

- A divergence $D(\theta : \theta')$ on a manifold M can induce a pair of conjugate connections (∇, ∇^*).
- The metric tensor g derived from D is given by:

$$g_{ij} = -\left.\frac{\partial^2 D(\theta : \theta')}{\partial\theta^i \partial\theta'^j}\right|_{\theta=\theta'} \tag{7}$$

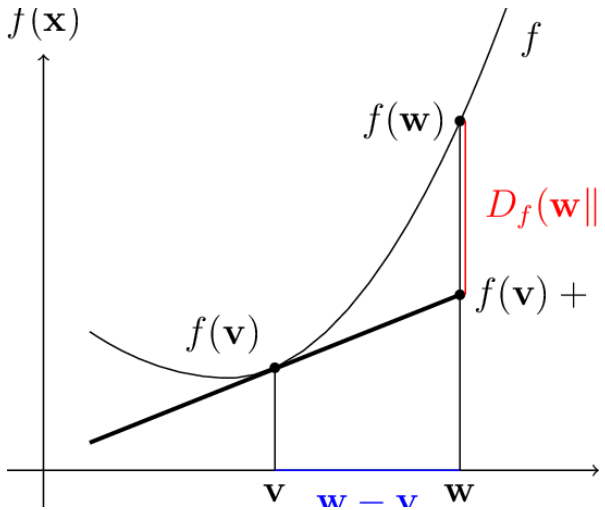
- The Christoffel symbols for ∇ and ∇^* are:

$$\Gamma^k_{ij} = -\left.\frac{\partial^3 D(\theta : \theta')}{\partial\theta^i \partial\theta^j \partial\theta'^k}\right|_{\theta=\theta'} \tag{8}$$

$$\Gamma^{*k}_{ij} = -\left.\frac{\partial^3 D(\theta : \theta')}{\partial\theta'^i \partial\theta'^j \partial\theta^k}\right|_{\theta=\theta'} \tag{9}$$

Bregman divergences

Consider a strictly convex smooth function $F(\theta)$ called a potential function
using this potential function we can derive the bregman divergence as follows.
which can be visualised as the difference between the potential at point θ' and the linear approximation of θ at point θ'.



$$B_F(\theta, \theta') = F(\theta) - F(\theta') - (\theta - \theta')^T \nabla F(\theta')$$

Bregman divergences

Examples

- Quadratic-From potential:
  $F(x) = \frac{1}{2}x^T Q x \Rightarrow B_F(\theta : \theta') = \frac{1}{2}(\theta - \theta')^T Q (\theta - \theta')$
- negative entropy potential:
  $F(p) = \sum_i p_i \log(p_i) \Rightarrow B_F(p : q) = D_{KL}(p||q)$
- free energy potential:
  $\mathcal{E} = \{p_\theta(x) = \exp\left(\sum_i t_i(x)\theta_i - F(\theta) + k(x)\right) | \theta \in \Theta\} \Rightarrow$
  $B_F(\theta : \theta') = D_{KL}(p_\theta||p_{\theta'})$ (to be revisited later)

Bregman divergences

Bregman geometry

Bregman divergences induce a special kind of information-geometric structure

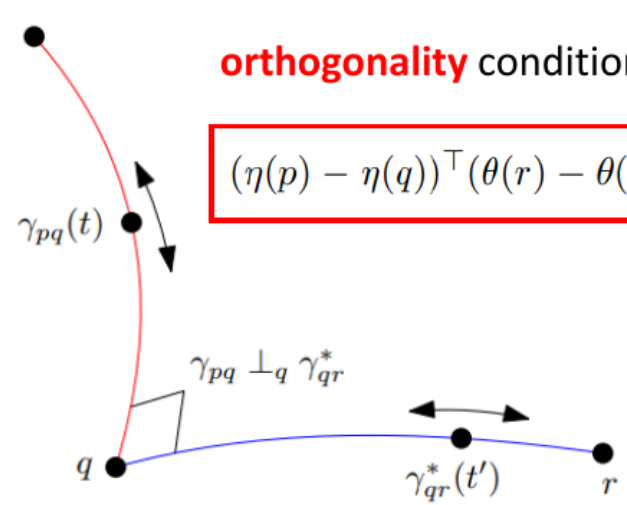
$$
\begin{aligned}
g^F &= \nabla^2 F(\theta) \\
\Gamma^F &= 0 \Rightarrow \nabla^F\text{-flat} \\
C^F_{ijk} &= \partial_i \partial_j \partial_k F(\theta)
\end{aligned}
$$

the convex conjugate also yields some usefull insights, ie the dual coordinates.

$$
F^*(\eta) = \sup_\theta \left\{ \theta^T \eta - F(\theta) \right\}, \quad \eta = \nabla F(\theta)
$$

Generalized Pythagorean theorem

Generalized Pythagoras' theorem

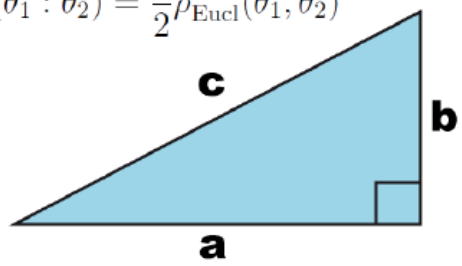


$$D_F(\gamma_{pq}(t) : \gamma_{qr}(t')) = D_F(\gamma_{pq}(t) : q) + D_F(q : \gamma^*_{qr}(t')), \quad \forall t, t' \in (0, 1).$$

Pythagoras' theorem in
the Euclidian geometry
Self-dual

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

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

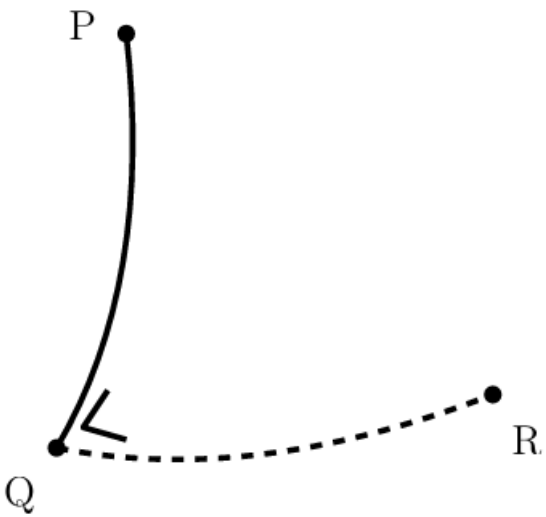


this is just an intuitive property of Bregman divergences!

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^T(\nabla F(\theta_2) - \nabla F(\theta_3))$$

Generalized Pythagorean theorem

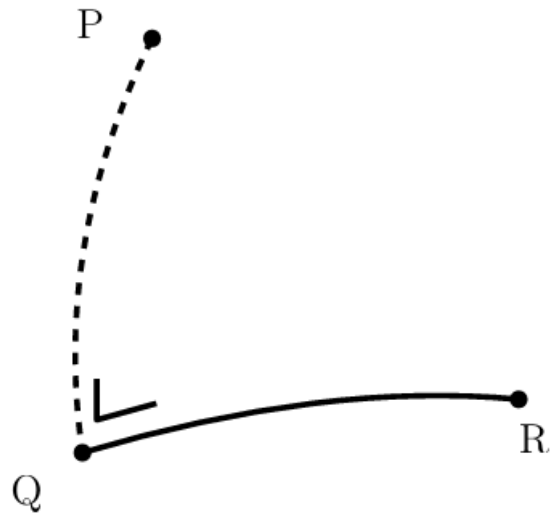
$\gamma^*(P,Q) \perp_F \gamma(Q,R)$



$$D(P:R) = D(P:Q) + D(Q:R)$$

$$B_F(\theta(P):\theta(R)) = B_F(\theta(P):\theta(Q)) + B_F(\theta(Q):\theta(R))$$

$\gamma(P,Q) \perp_F \gamma^*(Q,R)$



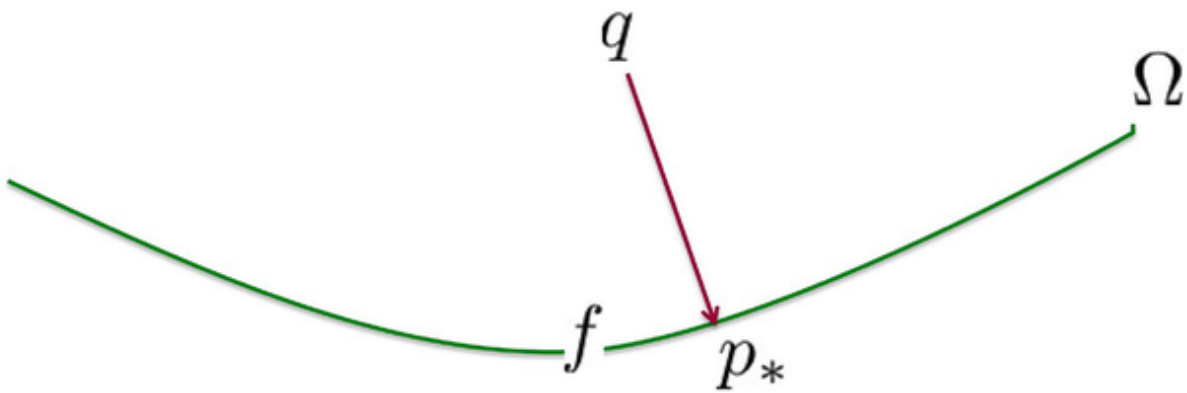
$$D^*(P:R) = D^*(P:Q) + D^*(Q:R)$$

$$B_{F^*}(\eta(P):\eta(R)) = B_{F^*}(\eta(P):\eta(Q)) + B_{F^*}(\eta(Q):\eta(R))$$

the case is particularly interesting if we consider all R from a particular $\nabla-$Flat or ∇^*-Flat submanifold.

Generalized Pythagorean theorem

Information projection uniqueness theorems



let p^* be the unique point that minimizes the Bregman divergence between p and some submanifold Ω. by the pythagorean theorem and positivity of divergences, this point will be unique.

Generalized Pythagorean theorem

Information projection uniqueness theorems

Some useful definitions

- $\nabla - projection: \quad P_S = \arg\min_Q D(\theta(P) : \theta(Q))$
- $\nabla^* - projection: \quad P_S^* = \arg\min_Q D(\theta(Q) : \theta(P))$
- $D(\mathcal{P} : \mathcal{Q}) = \min_{p \in \mathcal{P}, q \in \mathcal{Q}} D(p : q)$
- these will be revisited in MLE estimation, and the EM algorithm!

As we have seen before, the distance between two near points act χ^2-like, and in effect act quadratic.

$$D(P_\theta, p_{\theta+\delta}) = \delta^T I(\theta)\delta$$

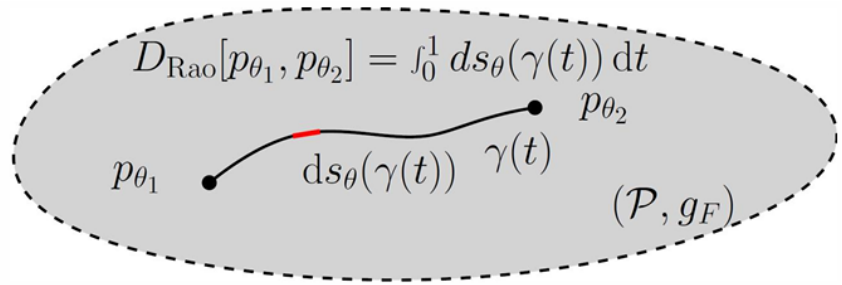
we can extend this idea and use the fisher information matrix as our metric, inducing the Fisher-Rao manifold.

$$\begin{aligned} g_F(u,v) &= \mathbb{E}[u(x)v(x)] = \mathrm{Cov}(u(x), v(x)) \\ g_F(\partial_i, \partial_j) &= \mathbb{E}_\theta[\partial_i l_x(\theta)\partial_j l_x(\theta)] = I_{ij}(\theta) \\ g_F(u,v) &= [u]_B^T I(\theta)[v]_B \end{aligned}$$

historically, information geometry started from this very idea!

Rao distance

Finding Fisher-Rao geodesics is a non-trivial tasks: No-known closed-form for the Fisher-Rao geodesic/distance between multivariate Gaussians



the length of the geodesic connecting the two points. an invariant distance metric.

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

$$\gamma(0) = \theta_1, \gamma(1) = \theta_2, \quad ds_\theta^2 = g_{ij}dx^i dx^j$$

Projections

e-Projections, m-Projections, ...

Exponential family, mixture family

- The exponential family $\mathcal{E} = \left\{p_\theta(x) = \exp(x^T\theta - F(\theta)\right\}$
- The mixture family
  $\mathcal{M} = \left\{p_\theta(x) = \sum_{i=1}^{D} \theta_i p_i(x) + (1 - \sum_{i=1}^{D} \theta_i)p_0(x)\right\}$

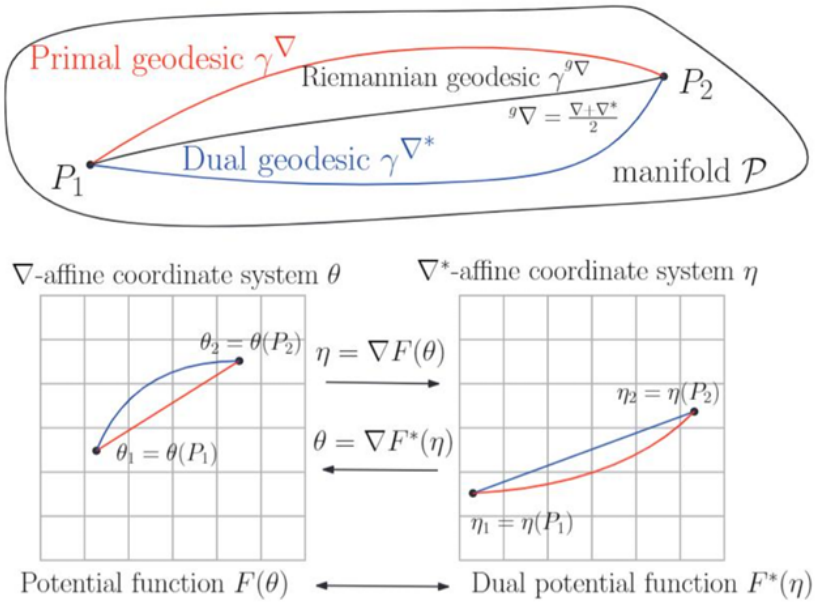
some very useful facts:

$$\nabla^e = (\nabla^m)^*, \quad \nabla^m = (\nabla^e)^*$$
$$\Gamma^e_{\mathcal{E}} = \Gamma^m_{\mathcal{E}} = \Gamma^e_{\mathcal{M}} = \Gamma^m_{\mathcal{M}} = 0$$

For the exponential family:
$F^*(\eta) = \sup_\theta \left\{\theta^T\eta - F(\theta)\right\}, \quad \eta = \nabla F(\theta) = \mathbb{E}[t(X)]$
Moment or mean parametrization: $p_\theta(x) = p^\eta(x)$

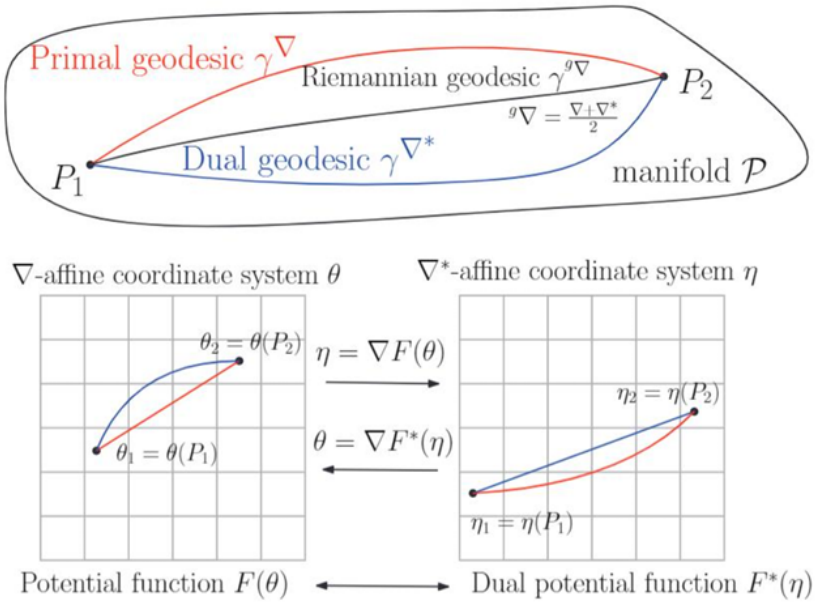
Hessian structures

- Primal geodesic:
  $\gamma_{p_{\theta_1} p_{\theta_2}}(t) = p_{(1-t)\theta_1 + t\theta_2}$
- Dual geodesic:
  $\gamma^*_{p_{\theta_1} p_{\theta_2}}(t) = p^{(1-t)\eta_1 + t\eta_2}$

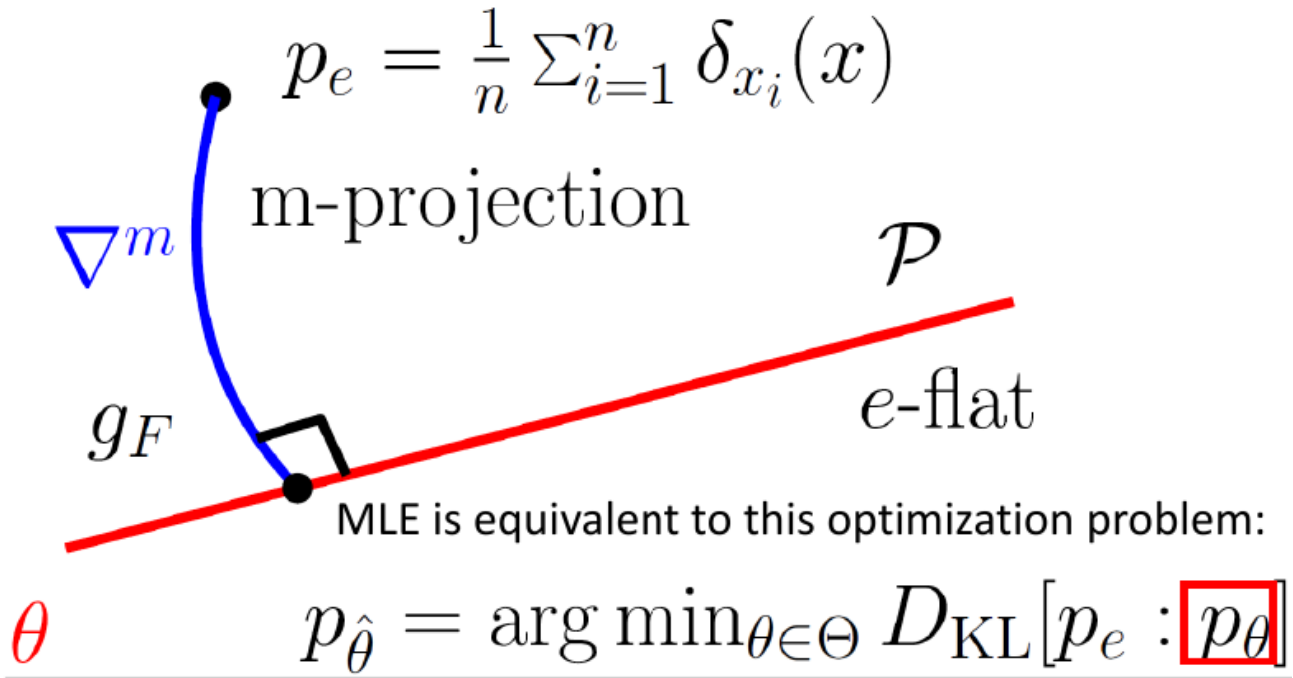


Hessian structures

- expectations are invariant to mixture families $\Rightarrow$ observations lead to m-flat structures.
- most distributions belong to the exponential family $\Rightarrow$ finding parameters is finding a point on e-flat structures.

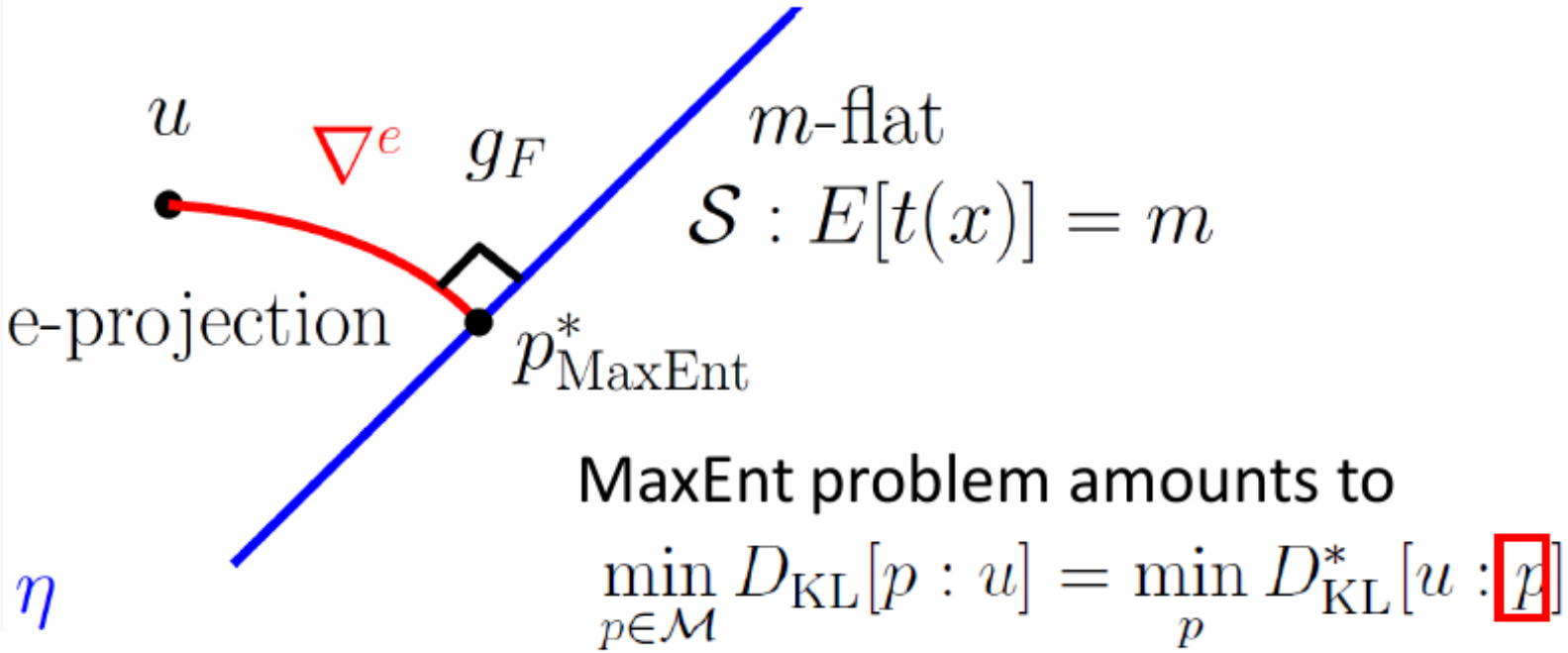


Maximum likelihood



$$p_{\hat{\theta}}^{MLE} = \mathrm{Proj}_{\mathcal{P}}^{\nabla^m}(p_e), \quad D_{\nabla^m} = D_{KL}$$

Maximum entropy

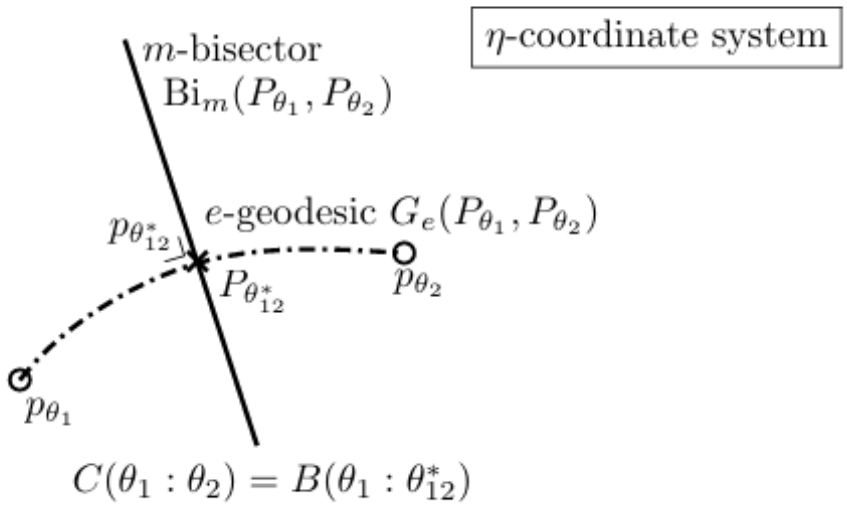


given k observations, $\mathbb{E}[m_i(x)] = a_i$, we will lie on an m-flat submanifold $\mathcal{M}_{n-k}$, after that the maximum entropy point will be the closest to the global maximum entropy point(uniform).

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

Chernoff information

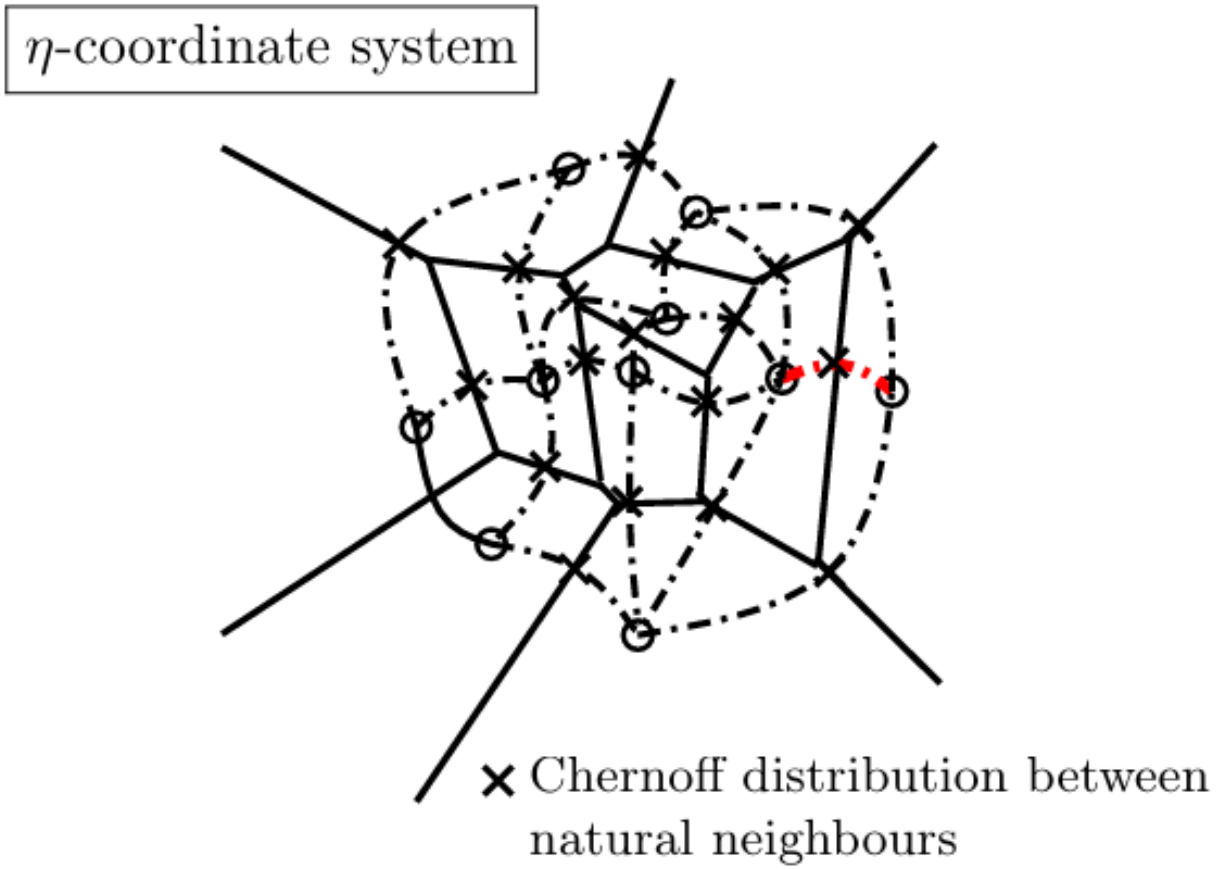
as many have seen before,
in Bayesian hypothesis
testing $P_e^n = 2^{-nC(P_1,P_2)}$
where
this point coincides with
the intersection of the
e-geodesic of the two points
and their m-bisector!



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

Chernoff information

Multiple hypthesis testing



Applications, and experiments

Intuition

- In the previous section, we saw a lot of new intuitions and ways to think about before known algorithms and methods.
- such methods such as e-projections and m-projections also hold valuable insight for stuff such as the EM-Algorithm, and NES-Algorithms.
- they can help us understand Restricted Boltzmann machines (RBMs)
- in signal processing they can help us with Principal Component Analysis (PCA), Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF),
- Game Theory, Mathematical Programming, ...

Natural Gradient Descent

One of the new methods, stemming straight from information geometry is the natural gradient descent.
as we stated before, we would like it so that all our methods, and how we talk about statistical phenomenon are invariant to how it is proposed to us, for example the coordinate system should not matter.
now let's analyse the popular gradient descent algorithm.

Natural Gradient Descent

Ordinary gradient descent (GD) method for minimizing a loss function $L(.)$:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta L_\theta(\theta_t)$$

- Depends on parametrization
- Plateu near singularities (almost degenerate Fisher information)

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

if we parametrize using $L_\eta(\eta) = L_\theta(\eta(\theta))$, we will most likely see a different series of η_t on our optimization, and on a non-convex loss function possibly different stationary points.

Natural Gradient Descent

Natural gradient is invariant to reparameterization and avoids plateaus:

$$\theta_{t+1} = \theta_t - \alpha \nabla^{\mathrm{NG}} L_\theta(\theta_t), \quad \nabla^{\mathrm{NG}} L_\theta(\theta) = g^{-1}(\theta) \nabla L_\theta(\theta)$$

in a sense, the natural gradiant is the Riemannian steepest descent, as in the *actual* steepest descent direction.
this is the same formulation as computing the gradient in a curved parametrization, i.e. spherical, ...

$$\begin{aligned} \eta = & \eta(\theta) \Rightarrow \nabla^{NG} L_\theta(\theta) = g^{-1}(\theta) \nabla_\theta L_\theta(\theta) = g^{-1}(\theta) \nabla_\theta L_\eta(\eta(\theta)) \\ & = g^{-1}(\theta)(\nabla_\theta(\eta))(\nabla_\eta L_\eta(\eta)) = g^{-1}(\eta) \nabla_\eta L_\eta(\eta) = \nabla^{\mathrm{NG}} L_\eta(\eta) \end{aligned}$$

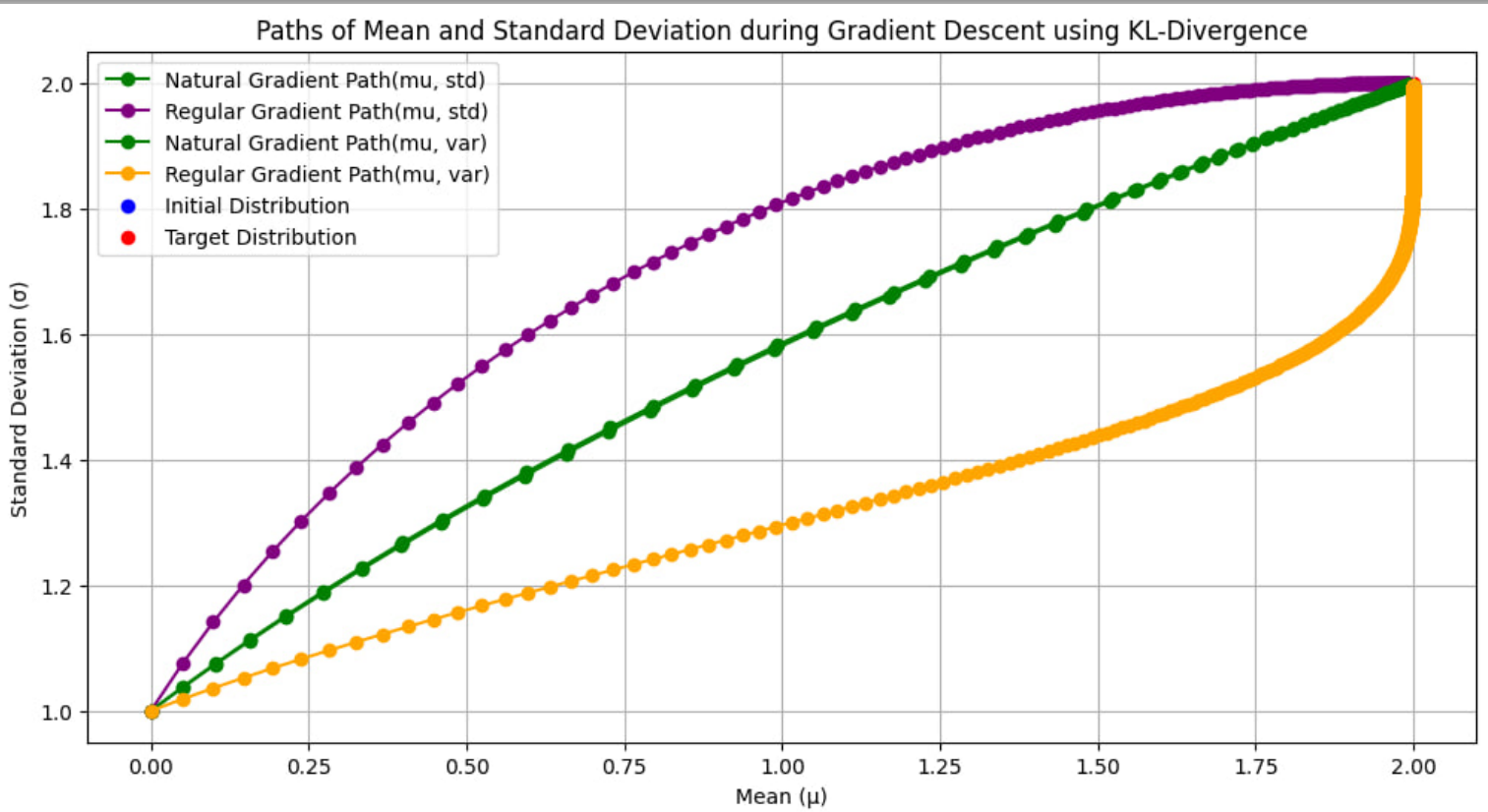
the natural gradient is invariant to parametrization!

Experimental results

In this section, we implemented the natural gradient algorithm on the problem of interpolating between two gaussian distributions smoothly.

Experimental results

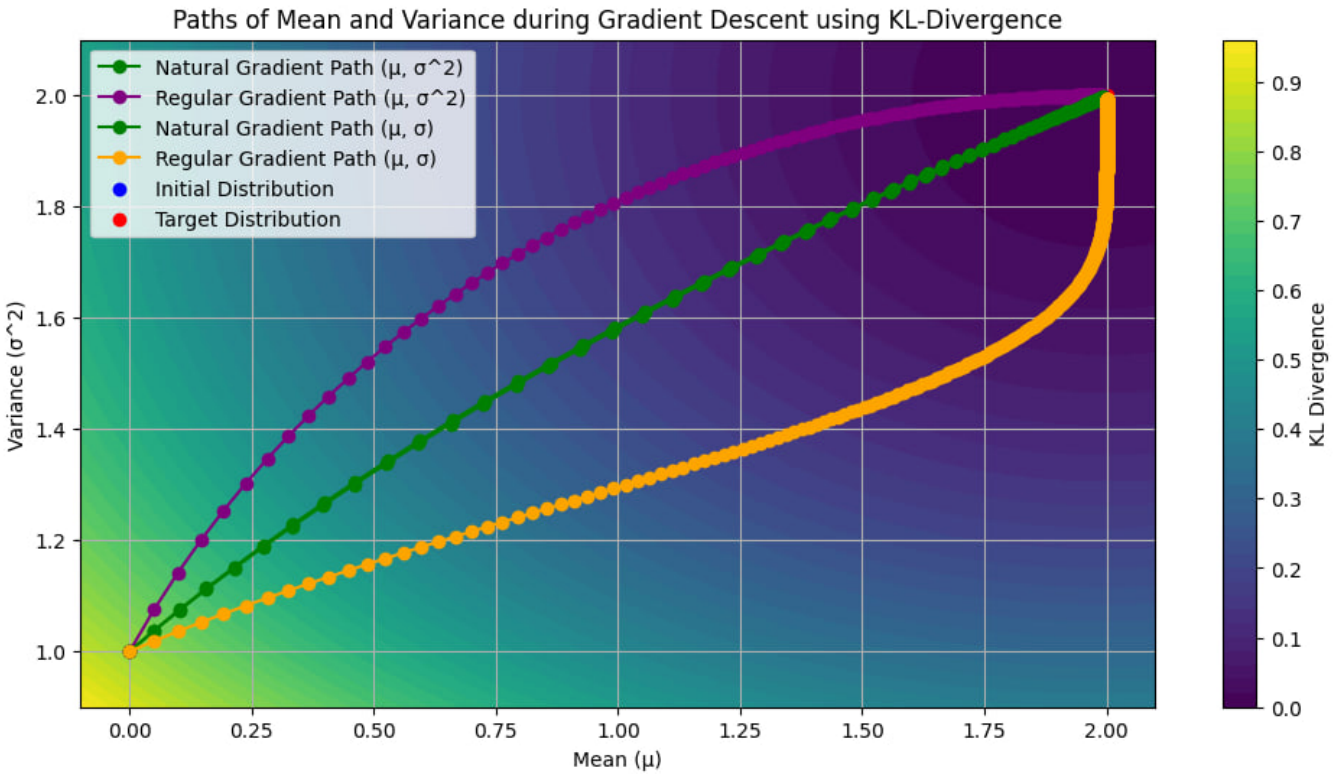
Natural gradients



$$L_\theta(\theta) = D_{KL}(\mathcal{N}(\theta_1^0, \theta_2^0)||\mathcal{N}(\theta_1^1, \theta_2^1)) \text{ for fixed } \theta^0$$

Experimental results

Natural gradients

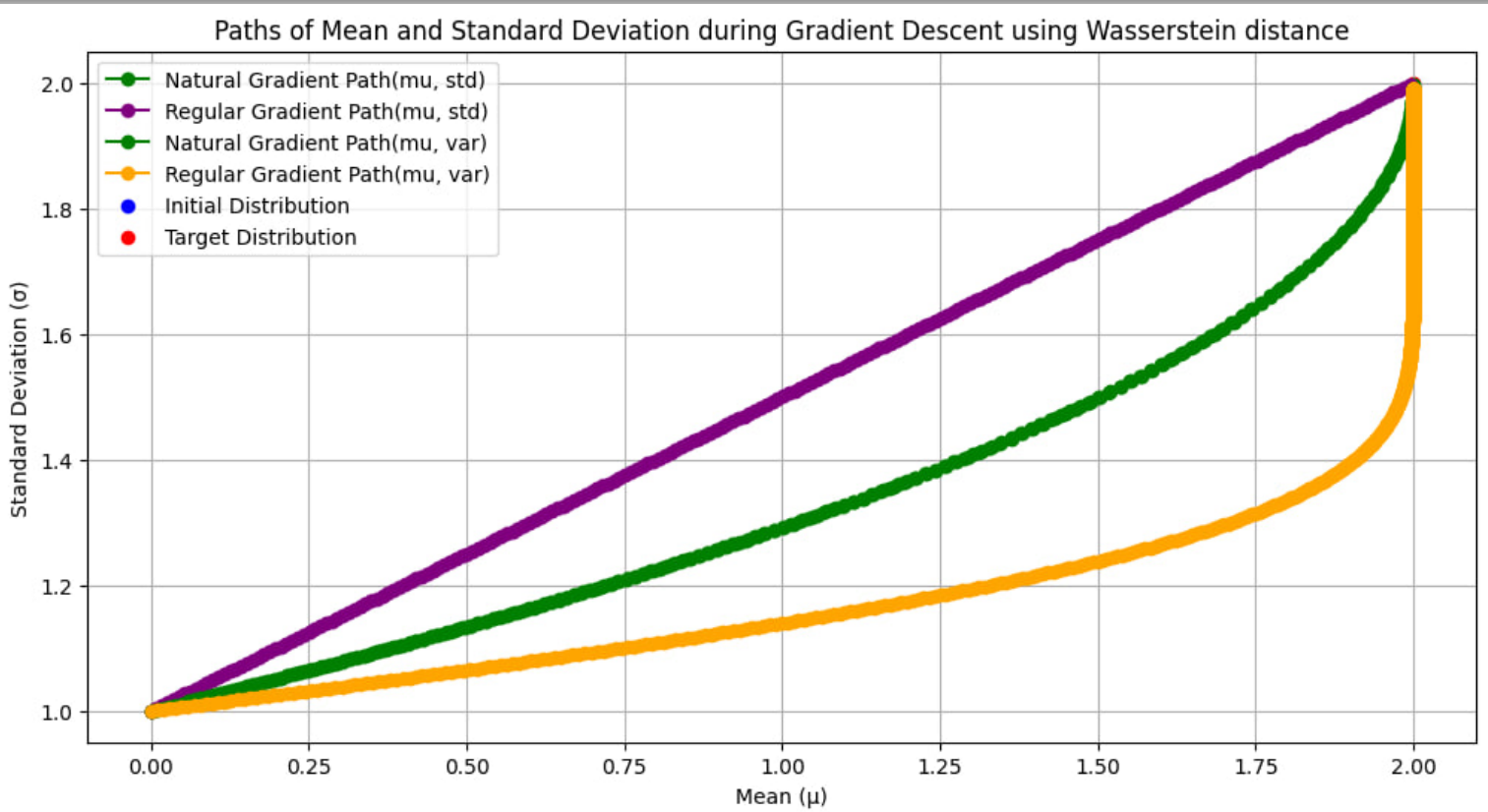


NGD: 67-iterations

GD-(μ, σ): 214-iterations, GD-(μ, σ^2): 1288-iterations

Experimental results

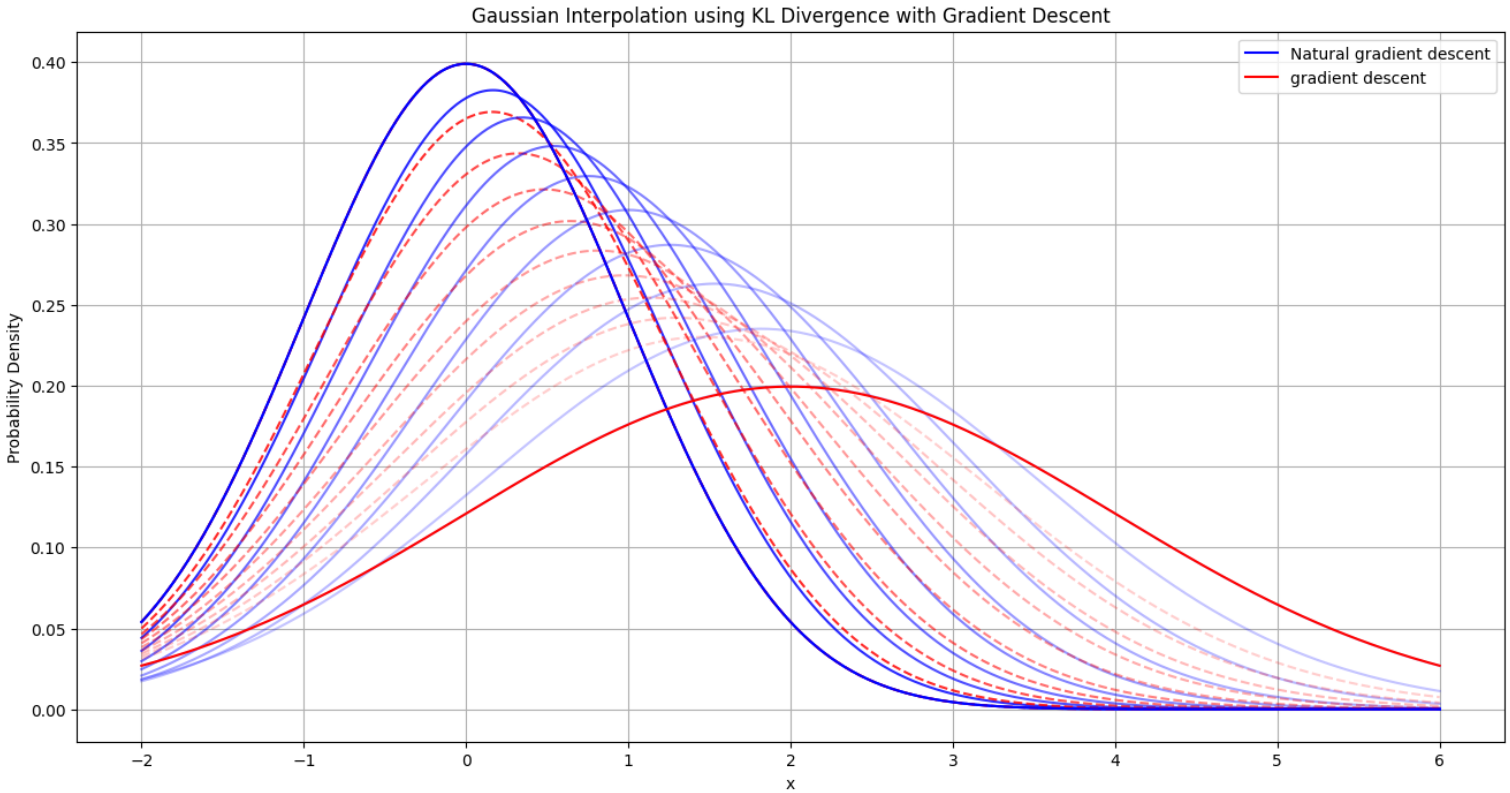
Natural gradients



$$L_\theta(\theta) = W(\mathcal{N}(\theta_1^0, \theta_2^0), \mathcal{N}(\theta_1^1, \theta_2^1)) \text{ for fixed } \theta^0$$

Experimental results

Interpolation between distributions



Transport using natural gradient descent and the wasserstein metric.

Q&A

Any Questions?

The End