Information geometry and symmetric spaces

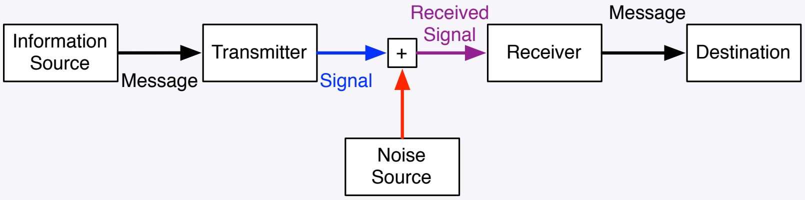
WOLFGANG GLOBKE

universität wien

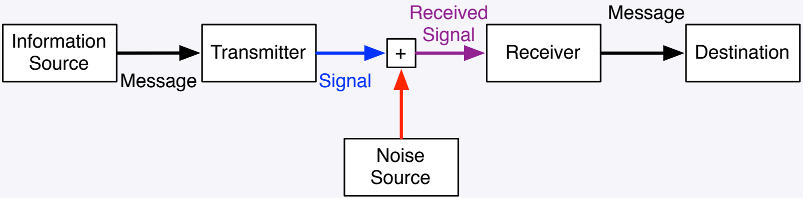
Colloquium
Centro de Investigación en Matemáticas, February 2019

Entropy and information

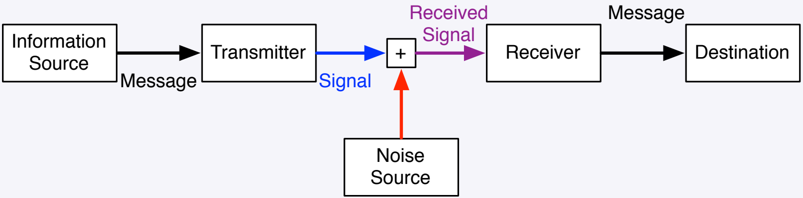
Communication model



Communication model



- Source alphabet $\mathcal{A} = \{a_1, \ldots, a_n\}$.
- Message space $\mathcal{A}^* = \{x_1 x_2 x_3 x_4 \ldots \mid x_i \in \mathcal{A}\}$.



- Source alphabet $\mathcal{A} = \{a_1, \dots, a_n\}$.
- Message space $\mathcal{A}^* = \{x_1 x_2 x_3 x_4 \dots \mid x_i \in \mathcal{A}\}$.
- Code alphabet $\mathcal{C} = \{c_1, \dots, c_b\}$, code space $\mathcal{C}^*$.
- Look for good encodings $\mathcal{A}^* \to \mathcal{C}^*$ to minimize noise effect and data transfer.

In his 1948 paper "The mathematical theory of communication",
Claude E. Shannon suggested a by now widely adopted measures of information.

In his 1948 paper "The mathematical theory of communication",
Claude E. Shannon suggested a by now widely adopted measures of information.

- $a \in \mathcal{A}$ appears with probability $p(a)$.

In his 1948 paper "The mathematical theory of communication",
Claude E. Shannon suggested a by now widely adopted measures of information.

- $a \in \mathcal{A}$ appears with probability $p(a)$.
- Information content of a should be $\log_b p(a)^{-1}$:

In his 1948 paper "The mathematical theory of communication", Claude E. Shannon suggested a by now widely adopted measures of information.

- $a \in \mathcal{A}$ appears with probability $p(a)$.
- Information content of a should be $\log_b p(a)^{-1}$:
  - Intuitively, we measure by linear comparison...
  - but many engineering parameters vary exponentially.

In his 1948 paper "The mathematical theory of communication", Claude E. Shannon suggested a by now widely adopted measures of information.

- $a \in \mathcal{A}$ appears with probability $p(a)$.
- Information content of a should be $\log_b p(a)^{-1}$:
  - Intuitively, we measure by linear comparison…
  - but many engineering parameters vary exponentially.
  - Example:
    Increasing a bit-wise ($b = 2$) representation by one bit *doubles* the number of possibilities, but increases the information by *one*.

In his 1948 paper "The mathematical theory of communication",
Claude E. Shannon suggested a by now widely adopted measures of information.

- $a \in \mathcal{A}$ appears with probability $p(a)$.
- Information content of a should be $\log_b p(a)^{-1}$:
  - Intuitively, we measure by linear comparison...
  - but many engineering parameters vary exponentially.
  - Example:
    Increasing a bit-wise ($b = 2$) representation by one bit *doubles* the number of possibilities, but increases the information by *one*.

Warning!
Intuitively, we associate some notion of "meaning" with "information".
But semantic aspects are irrelevant for the engineering problem!

If a process produces symbols $a \in \mathcal{A}$ with probabilities $p(a)$,
can we assign an information to this process?

Shannon's entropy

If a process produces symbols $a \in \mathcal{A}$ with probabilities $p(a)$,
can we assign an information to this process?

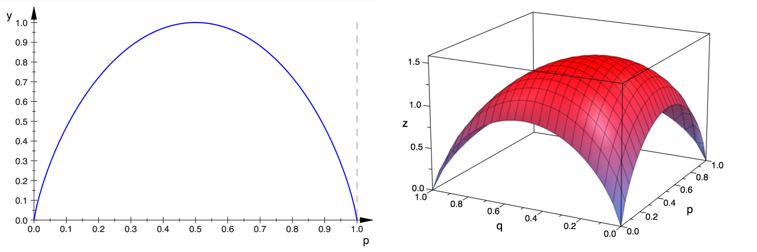
The entropy of a random variable X with values in $\mathcal{A}$ is the expected information,

$$H(X) = -\sum_{a \in \mathcal{A}} p(a) \log p(a).$$

If a process produces symbols $a \in \mathcal{A}$ with probabilities $p(a)$,
can we assign an information to this process?

The entropy of a random variable X with values in $\mathcal{A}$ is the expected information,

$$H(X) = -\sum_{a \in \mathcal{A}} p(a) \log p(a).$$



Optimal codelength

If H is the entropy of the symbols $\mathcal{A}$, how efficient can an encoding $\mathsf{c} : \mathcal{A} \to \mathcal{C}^*$ be?

If H is the entropy of the symbols $\mathcal{A}$, how efficient can an encoding $\mathsf{c} : \mathcal{A} \to \mathcal{C}^*$ be?

The expected codelength is

$$L = \sum_{a_i \in \mathcal{A}} p(a_i)\text{length}(\mathsf{c}(a_i)).$$

If H is the entropy of the symbols $\mathcal{A}$, how efficient can an encoding $\mathsf{c} : \mathcal{A} \to \mathcal{C}^*$ be?

The expected codelength is

$$L = \sum_{a_i \in \mathcal{A}} p(a_i)\text{length}(\mathsf{c}(a_i)).$$

Shannon's Source Coding Theorem
If $\mathcal{C}^*$ is a prefix code (no codeword prefix of another), then

$$L \geq H(X).$$

If H is the entropy of the symbols $\mathcal{A}$, how efficient can an encoding $\mathsf{c} : \mathcal{A} \to \mathcal{C}^*$ be?

The expected codelength is

$$L = \sum_{a_i \in \mathcal{A}} p(a_i)\text{length}(\mathsf{c}(a_i)).$$

Shannon's Source Coding Theorem
If $\mathcal{C}^*$ is a prefix code (no codeword prefix of another), then

$$L \geq H(X).$$

How close can real codes get?

The Huffman code (1952) realizes

$$H(X) \;\leq\; L \;<\; H(X) + 1.$$

The Huffman code (1952) realizes

$$H(X) \;\leq\; L \;<\; H(X) + 1.$$

This allows us to interpret the entropy (base 2) as

$$H(X) \approx \begin{array}{l}\text{expected number of (clever) Yes/No-questions} \\ \text{to determine which } a_i \in \mathcal{A} \text{ was received.}\end{array}$$

The Huffman code (1952) realizes

$$H(X) \;\leq\; L \;<\; H(X) + 1.$$

This allows us to interpret the entropy (base 2) as

$$H(X) \approx \begin{array}{l}\text{expected number of (clever) Yes/No-questions} \\ \text{to determine which } a_i \in \mathcal{A} \text{ was received.}\end{array}$$

"Proof"

- For any conceivable sequence of Yes/No-question,
  each question can be interpreted as one bit in an encoding of $\mathcal{A}$.
- Huffman code provides a "clever" sequence of questions.

Suppose we have two possible probability distributions p, q for X.

Suppose we have two possible probability distributions p, q for X.

The divergence (a.k.a. Kullback-Leibler distance a.k.a. relative entropy) of X is

$$D(p\|q) = \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)} = \mathsf{E}_p(\log p - \log q).$$

Suppose we have two possible probability distributions p, q for X.

The divergence (a.k.a. Kullback-Leibler distance a.k.a. relative entropy) of X is

$$D(p\|q) = \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)} = \mathsf{E}_p(\log p - \log q).$$

It is used as a "distance measure" for probability distributions.

Suppose we have two possible probability distributions p, q for X.

The divergence (a.k.a. Kullback-Leibler distance a.k.a. relative entropy) of X is

$$D(p\|q) = \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)} = \mathsf{E}_p(\log p - \log q).$$

It is used as a “distance measure” for probability distributions.

Properties

1. $D(p\|q) \geq 0$.
2. $D(p\|q) = 0 \Leftrightarrow p = q$.

Suppose we have two possible probability distributions p, q for X.

The divergence (a.k.a. Kullback-Leibler distance a.k.a. relative entropy) of X is

$$D(p\|q) = \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)} = \mathsf{E}_p(\log p - \log q).$$

It is used as a "distance measure" for probability distributions.

Properties

1. $D(p\|q) \geq 0$.
2. $D(p\|q) = 0 \Leftrightarrow p = q$.
3. $D(p\|q) \neq D(q\|p)$ in general.

Suppose we have two possible probability distributions p, q for X.

The divergence (a.k.a. Kullback-Leibler distance a.k.a. relative entropy) of X is

$$D(p\|q) = \sum_{a\in\mathcal{A}} p(a)\log\frac{p(a)}{q(a)} = \mathsf{E}_p(\log p - \log q).$$

It is used as a “distance measure” for probability distributions.

Properties

1. $D(p\|q) \geq 0$.
2. $D(p\|q) = 0 \Leftrightarrow p = q$.
3. $D(p\|q) \neq D(q\|p)$ in general.
4. D does not satisfy the triangle inequality.

Suppose we are using the wrong distribution q (instead of the correct one p) in Shannon's Source Coding Theorem:

Suppose we are using the wrong distribution q (instead of the correct one p) in Shannon's Source Coding Theorem:

$$H(p) + D(p\|q) \;\leq\; L_{\boldsymbol{p}} \;<\; H(p) + D(p\|q) + 1$$

Suppose we are using the wrong distribution q (instead of the correct one p) in Shannon's Source Coding Theorem:

$$H(p) + D(p\|q) \ \leq\ L_p \ <\ H(p) + D(p\|q) + 1$$

where L_p is the expected length (under p) of a code constructed under the assumption of q.

Suppose we are using the wrong distribution q (instead of the correct one p) in Shannon's Source Coding Theorem:

$$H(p) + D(p\|q) \;\leq\; L_p \;<\; H(p) + D(p\|q) + 1$$

where L_p is the expected length (under p) of a code constructed under the assumption of q.

$D(p\|q)$ makes similar appearances in many other identies in information theory.

Information and statistics

Most probability distributions in statistics depend on a finite number of parameters $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

Most probability distributions in statistics depend on a finite number of parameters $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

Examples

- The normal distribution $N(\mu, \sigma)$ depends on mean $\theta_1 = \mu$ and variance $\theta_2 = \sigma^2$.
- A distribution on a finite set $\Omega = \{x_1, \ldots, x_k\}$ depends on parameters $\theta_1 = p(x_1), \ldots, \theta_{k-1} = p(x_{k-1})$.

Most probability distributions in statistics depend on a finite number of parameters $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

Examples

- The normal distribution $N(\mu, \sigma)$ depends on mean $\theta_1 = \mu$ and variance $\theta_2 = \sigma^2$.
- A distribution on a finite set $\Omega = \{x_1, \ldots, x_k\}$ depends on parameters $\theta_1 = p(x_1), \ldots, \theta_{k-1} = p(x_{k-1})$.

Standard problem

- Assume data is distributed according to a certain type of distribution $p(x \mid \theta)$ on a sample space Ω.
- Task: Estimate $\theta \in \Theta$ from observed data $\mathtt{y}_1, \ldots, \mathtt{y}_d \in \Omega$.
- An estimator $\hat{\theta}$ for θ is a function $\hat{\theta} : \Omega^d \to \Theta$.

Example: Maximum likelihood estimator

If $\mathtt{y}_1, \ldots, \mathtt{y}_d$ are independent observations, the likelihood of parameter θ is

$$L(\theta \mid \mathtt{y}_1, \ldots, \mathtt{y}_d) = \prod_{i=1}^{d} p(\mathtt{y}_i \mid \theta).$$

Example: Maximum likelihood estimator

If $\mathtt{y}_1, \ldots, \mathtt{y}_d$ are independent observations, the likelihood of parameter θ is

$$L(\theta \mid \mathtt{y}_1, \ldots, \mathtt{y}_d) = \prod_{i=1}^{d} p(\mathtt{y}_i \mid \theta).$$

More convenient: the log-likelihood (same maxima as likelihood)

$$\log L(\theta \mid \mathtt{y}_1, \ldots, \mathtt{y}_d) = \sum_{i=1}^{d} \log p(\mathtt{y}_i \mid \theta).$$

Example: Maximum likelihood estimator

If $\mathtt{y}_1, \ldots, \mathtt{y}_d$ are independent observations, the likelihood of parameter θ is

$$L(\theta \mid \mathtt{y}_1, \ldots, \mathtt{y}_d) = \prod_{i=1}^{d} p(\mathtt{y}_i \mid \theta).$$

More convenient: the log-likelihood (same maxima as likelihood)

$$\log L(\theta \mid \mathtt{y}_1, \ldots, \mathtt{y}_d) = \sum_{i=1}^{d} \log p(\mathtt{y}_i \mid \theta).$$

The maximum likelihood estimator $\hat{\theta}$ is found by maximizing $\log L$; solve for $\hat{\theta}$:

$$\mathrm{grad}_\theta \log L = \frac{1}{L}\, \mathrm{grad}_\theta\, L = 0.$$

Heuristics

The second derivative

$$\mathrm{Hess}_\theta \log L = \left(\frac{\partial^2 \log L}{\partial \theta_i \, \partial \theta_j} \right)_{\theta = \hat{\theta}}$$

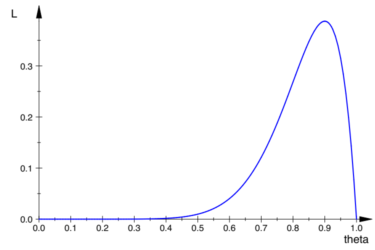
determines the curvature of $\log L$ at $\theta = \hat{\theta}$.

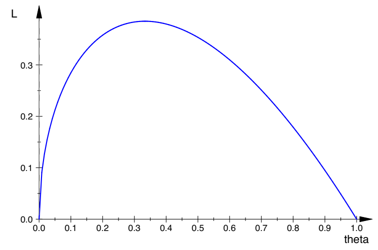
Heuristics

The second derivative

$$\mathrm{Hess}_\theta \log L = \left(\frac{\partial^2 \log L}{\partial\theta_i \, \partial\theta_j} \right)_{\theta=\hat{\theta}}$$

determines the curvature of $\log L$ at $\theta = \hat{\theta}$.



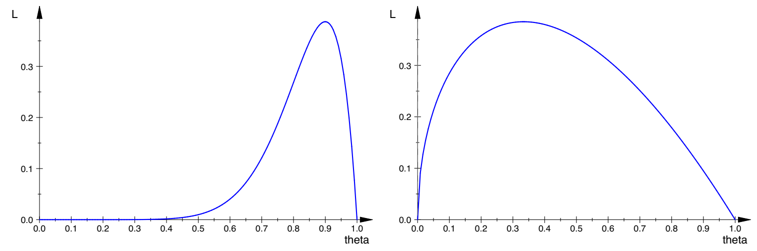


Heuristics

The second derivative

$$\mathrm{Hess}_\theta \log L = \left(\frac{\partial^2 \log L}{\partial\theta_i \, \partial\theta_j} \right)_{\theta=\hat{\theta}}$$

determines the curvature of $\log L$ at $\theta = \hat{\theta}$.



The larger the curvature, the more precise the estimator is.

In “On the Mathematical Foundations of Theoretical Statistics” in 1921, Ronald A. Fisher introduced a different concept of information, which is supposed to describe the contribution of a parameter to a model.

For $\theta \in \Theta$, the Fisher information is

$$\mathrm{g}(\theta) = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta).$$

In "On the Mathematical Foundations of Theoretical Statistics" in 1921, Ronald A. Fisher introduced a different concept of information, which is supposed to describe the contribution of a parameter to a model.

For $\theta \in \Theta$, the Fisher information is

$$\mathrm{g}(\theta) = -\,\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta).$$

Fact

For i.i.d. random variables $X_1, \ldots, X_n$ with joint probability $p_n(X_1, \ldots, X_n \mid \theta)$,

$$\mathrm{g}_n(\theta) = n\,\mathrm{g}(\theta).$$

In "On the Mathematical Foundations of Theoretical Statistics" in 1921, Ronald A. Fisher introduced a different concept of information, which is supposed to describe the contribution of a parameter to a model.

For $\theta \in \Theta$, the Fisher information is

$$g(\theta) = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta).$$

Fact
For i.i.d. random variables $X_1, \ldots, X_n$ with joint probability $p_n(X_1, \ldots, X_n \mid \theta)$,

$$g_n(\theta) = n g(\theta).$$

Cramér-Rao inequality (1945)
The variance of any unbiased estimator (i.e. expected error from true value $= 0$*) has "lower bound"*

$$\mathsf{Var}_\theta(\hat{\theta}) \geq g(\theta)^{-1}$$

(meaning the $\mathsf{Var}_\theta(\hat{\theta}) - g(\theta)^{-1}$ *is positive semidefinite).*

Information geometry

How exactly does $g(\theta) = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta)$ determine the (average) curvature?

How exactly does $\mathrm{g}(\theta) = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta)$ determine the (average) curvature?

In 1945, C. Radhakrishna Rao observed that the parameter space Θ becomes a Riemannian manifold (M, g) with Fisher information $\mathrm{g}(\theta)$ as metric tensor at the point $\theta \in M$

How exactly does $g(\theta) = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta)$ determine the (average) curvature?

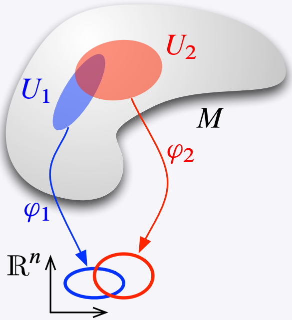
In 1945, C. Radhakrishna Rao observed that the parameter space Θ becomes a Riemannian manifold (M, g) with Fisher information $g(\theta)$ as metric tensor at the point $\theta \in M$ (assuming M and g are sufficiently "well-behaved", which they usually are).

A differentiable manifold M is a (suitable) topological space, covered by a family $\{(U, \varphi)\}$ (coordinate charts) of open sets U with homeomorphisms $\varphi : U \to \mathbb{R}^n$.

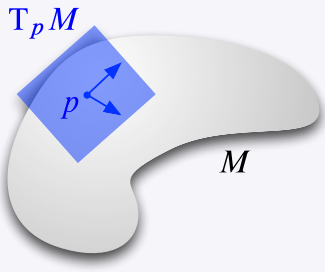
- Coordinate changes $\varphi_1 \circ \varphi_2^{-1}$ are C^∞-maps.
- $\dim M = n$.

Examples

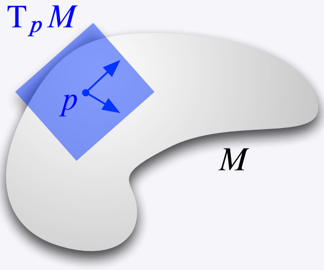
- $\mathbb{R}^n$ itself.
- n-Sphere $\mathbf{S}^n$.
- Torus $\mathbf{T}^n$.
- Matrix groups $GL_n(\mathbb{R})$, $SL_n(\mathbb{R})$, O_n.
- Well-behaved parameter spaces Θ in statistics.



A tangent vector at $p \in M$ is the equivalence class of all C^∞-curves $c : (-\varepsilon, \varepsilon) \to M$ with $c(0) = p$ and whose first derivatives (in charts) coincide.
The tangent space $T_p M$ at p is the space spanned by the tangent vectors at p.



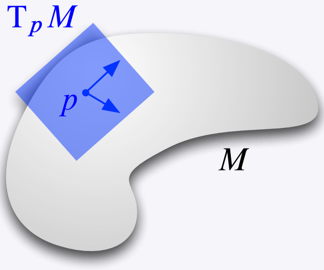
A tangent vector at $p \in M$ is the equivalence class of all C^∞-curves $c : (-\varepsilon, \varepsilon) \to M$ with $c(0) = p$ and whose first derivatives (in charts) coincide. The tangent space $T_p M$ at p is the space spanned by the tangent vectors at p.



A Riemannian manifold (M, g) is a manifold M with a family $g = (g_p)_{p \in M}$ of positive definite scalar products g_p in $T_p M$ (the Riemannian metric), and g_p depends differentiably on p.

Differential geometry dictionary (II)

A tangent vector at $p \in M$ is the equivalence class of all C^∞-curves $c : (-\varepsilon, \varepsilon) \to M$ with $c(0) = p$ and whose first derivatives (in charts) coincide.
The tangent space $T_p M$ at p is the space spanned by the tangent vectors at p.



A Riemannian manifold (M, g) is a manifold M with a family $g = (g_p)_{p \in M}$ of positive definite scalar products g_p in $T_p M$ (the Riemannian metric), and g_p depends differentiably on p.
This induces a metric on M by via $\mathrm{dist}(p, q) = \inf_\gamma \int_a^b \|\gamma'(t)\|_{\gamma(t)} \mathrm{d}t$.

Differential geometry dictionary (III)

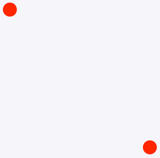
How to compare vectors in different tangent spaces $\mathrm{T}_p M$ and $\mathrm{T}_q M$?

Differential geometry dictionary (III)

How to compare vectors in different tangent spaces $\mathrm{T}_p M$ and $\mathrm{T}_q M$?

This is made possible by an affine connection (also covariant derivative) $\nabla_X Y$ of vector fields X, Y on M (“directional derivative”).

- ∇ defines a parallel transport along curves $c : (a, b) \to M$ by $\nabla_{c'(t)} X = 0$.
- “Straight lines” are given by geodesic curves, defind by $\nabla_{c'(t)} c'(t) = 0$.

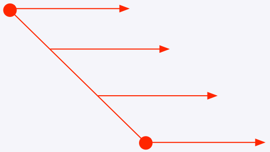


Differential geometry dictionary (III)

How to compare vectors in different tangent spaces $\mathrm{T}_p M$ and $\mathrm{T}_q M$?

This is made possible by an affine connection (also covariant derivative) $\nabla_X Y$ of vector fields X, Y on M (“directional derivative”).

- ∇ defines a parallel transport along curves $c : (a, b) \to M$ by $\nabla_{c'(t)} X = 0$.
- “Straight lines” are given by geodesic curves, defind by $\nabla_{c'(t)} c'(t) = 0$.

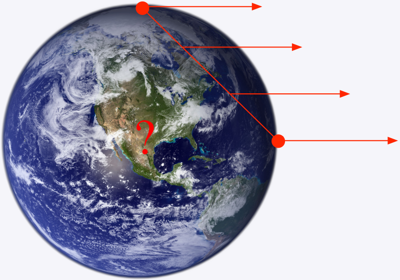


Differential geometry dictionary (III)

How to compare vectors in different tangent spaces $\mathrm{T}_p M$ and $\mathrm{T}_q M$?

This is made possible by an affine connection (also covariant derivative) $\nabla_X Y$ of vector fields X, Y on M ("directional derivative").

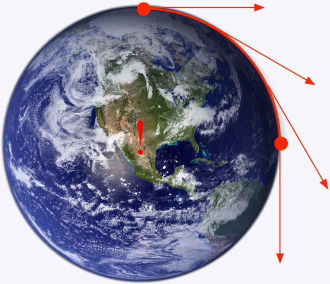
- ∇ defines a parallel transport along curves $c : (a, b) \to M$ by $\nabla_{c'(t)} X = 0$.
- "Straight lines" are given by geodesic curves, defind by $\nabla_{c'(t)} c'(t) = 0$.



How to compare vectors in different tangent spaces $\mathrm{T}_p M$ and $\mathrm{T}_q M$?

This is made possible by an affine connection (also covariant derivative) $\nabla_X Y$ of vector fields X, Y on M ("directional derivative").

- ∇ defines a parallel transport along curves $c : (a, b) \to M$ by $\nabla_{c'(t)} X = 0$.
- "Straight lines" are given by geodesic curves, defind by $\nabla_{c'(t)} c'(t) = 0$.



Differential geometry dictionary (IV)

Given M with a covariant derivative ∇, the curvature tensor R of ∇ is defined for vector fields X, Y, Z on M by

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

where $[X,Y]$ is the commutator of vector fields ($[X,Y] = X \circ Y - Y \circ X$ as differential operators).

Given M with a covariant derivative ∇, the curvature tensor R of ∇ is defined for vector fields X, Y, Z on M by

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

where $[X, Y]$ is the commutator of vector fields ($[X, Y] = X \circ Y - Y \circ X$ as differential operators).
We say M (or ∇) is flat if $\mathrm{R} = 0$ at all $p \in M$.

Differential geometry dictionary (IV)

Given M with a covariant derivative ∇, the curvature tensor R of ∇ is defined for vector fields X, Y, Z on M by

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

where $[X, Y]$ is the commutator of vector fields ($[X, Y] = X \circ Y - Y \circ X$ as differential operators).
We say M (or ∇) is flat if $\mathrm{R} = 0$ at all $p \in M$.

A Riemannian manifolds (M, g) has a canonical Levi-Civita connection ∇^{g} with

$$\nabla^{\mathrm{g}} \mathrm{g} = 0, \quad \nabla^{\mathrm{g}}_X Y - \nabla^{\mathrm{g}}_Y X = [X, Y].$$

Differential geometry dictionary (IV)

Given M with a covariant derivative ∇, the curvature tensor R of ∇ is defined for vector fields X, Y, Z on M by

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

where $[X,Y]$ is the commutator of vector fields ($[X,Y] = X \circ Y - Y \circ X$ as differential operators).
We say M (or ∇) is flat if $\mathrm{R} = 0$ at all $p \in M$.

A Riemannian manifolds (M, g) has a canonical Levi-Civita connection ∇^{g} with

$$\nabla^{\mathrm{g}} \mathrm{g} = 0, \quad \nabla^{\mathrm{g}}_X Y - \nabla^{\mathrm{g}}_Y X = [X,Y].$$

On a Riemannian manifold, the sectional curvature of tangent planes spanned by X_p, Y_p at $p \in M$, is

$$\mathrm{K}(X_p, Y_p) = \frac{\mathrm{g}(\mathrm{R}^{\mathrm{g}}(X_p, Y_p) X_p, Y_p)}{\mathrm{area}(X_p, Y_p)}.$$

A statistical manifold (M, g) is a manifold M of probability distributions $p(\cdot \mid \theta)$, with parameters $\theta = (\theta_1, \ldots, \theta_n)$ as coordinates, and g is the Fisher metric

$$\mathrm{g}_\theta = -\,\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta).$$

A statistical manifold (M, g) is a manifold M of probability distributions $p(\cdot \mid \theta)$, with parameters $\theta = (\theta_1, \ldots, \theta_n)$ as coordinates, and g is the Fisher metric

$$\mathrm{g}_\theta = -\mathsf{E}_\theta \operatorname{Hess}_\theta \log p(X \mid \theta).$$

Define an affine connection ∇^{I} on M by

$$\mathrm{g}_\theta(\nabla^{\mathrm{I}}_{X_i} X_j, X_k) = \mathsf{E}_\theta \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x \mid \theta) \frac{\partial}{\partial\theta_k} \log p(x \mid \theta) \right),$$

where $X_i = \frac{\partial}{\partial\theta_i}$ are the coordinate vector fields (in general $\nabla^{\mathrm{I}} \neq \nabla^{\mathrm{g}}$).

A statistical manifold (M, g) is a manifold M of probability distributions $p(\cdot \mid \theta)$, with parameters $\theta = (\theta_1, \ldots, \theta_n)$ as coordinates, and g is the Fisher metric

$$\mathrm{g}_\theta = -\,\mathsf{E}_\theta \,\mathrm{Hess}_\theta \log p(X \mid \theta).$$

Define an affine connection ∇^{I} on M by

$$\mathrm{g}_\theta(\nabla^{\mathrm{I}}_{X_i} X_j, X_k) = \mathsf{E}_\theta \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x \mid \theta) \frac{\partial}{\partial\theta_k} \log p(x \mid \theta) \right),$$

where $X_i = \frac{\partial}{\partial\theta_i}$ are the coordinate vector fields (in general $\nabla^{\mathrm{I}} \neq \nabla^{\mathrm{g}}$).

The relative entropy $D(p\|q)$ as a function of $q \in M$ has an expansion at the point $p \in M$

$$D(p\|q) = \frac{1}{2} \sum_{i,j} \mathrm{g}_p(X_i, X_j)\delta_i\delta_j + \frac{1}{6} \sum_{i,j,k} \left(\frac{\partial}{\partial\theta_i} \mathrm{g}_p(X_j, X_k) + \mathrm{g}_p(\nabla^{\mathrm{I}}_{X_j} X_k, X_i) \right) \delta_i\delta_j\delta_k$$

with $\delta_i = \theta_i(p) - \theta_i(q)$.

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \ldots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \ldots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

The dual coordinates $\eta_1, \ldots, \eta_n$ of θ are defined by

$$\mathrm{g}(X_i, Y_j) = \delta_i^j \quad \text{(Kronecker symbol)},$$

where X_i and Y_j are coordinate vector fields for θ_i and η_j.

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \ldots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

The dual coordinates $\eta_1, \ldots, \eta_n$ of θ are defined by

$$\mathrm{g}(X_i, Y_j) = \delta_i^j \quad \text{(Kronecker symbol)},$$

where X_i and Y_j are coordinate vector fields for θ_i and η_j. Then

$$\left.\frac{\partial \eta_j}{\partial \theta_i}\right|_p = \mathrm{g}_p(X_i, X_j)$$

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \ldots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

The dual coordinates $\eta_1, \ldots, \eta_n$ of θ are defined by

$$\mathrm{g}(X_i, Y_j) = \delta_i^j \quad \text{(Kronecker symbol)},$$

where X_i and Y_j are coordinate vector fields for θ_i and η_j. Then

$$\frac{\partial \eta_j}{\partial \theta_i}\Big|_p = \mathrm{g}_p(X_i, X_j) = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\Big|_{q=p} D(p\|q).$$

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \ldots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

The dual coordinates $\eta_1, \ldots, \eta_n$ of θ are defined by

$$\mathrm{g}(X_i, Y_j) = \delta_i^j \quad \text{(Kronecker symbol)},$$

where X_i and Y_j are coordinate vector fields for θ_i and η_j. Then

$$\frac{\partial \eta_j}{\partial \theta_i}\Big|_p = \mathrm{g}_p(X_i, X_j) = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\Big|_{q=p} D(p\|q).$$

The solution ψ of the differential equation

$$\frac{\partial \psi}{\partial \theta_i} = \eta_i$$

then satisfies

$$\mathrm{Hess}_\theta\, \psi = \mathrm{g}.$$

If ∇^{I} is flat, then near each point, M is equivalent to an open subset of $\mathbb{R}^n$.
We may then assume w.l.o.g. that $\theta_1, \dots, \theta_n$ are the canonical coordinates of $\mathbb{R}^n$.

The dual coordinates $\eta_1, \dots, \eta_n$ of θ are defined by

$$g(X_i, Y_j) = \delta_i^j \quad \text{(Kronecker symbol)},$$

where X_i and Y_j are coordinate vector fields for θ_i and η_j. Then

$$\frac{\partial \eta_j}{\partial \theta_i}\Big|_p = g_p(X_i, X_j) = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\Big|_{q=p} D(p\|q).$$

The solution ψ of the differential equation

$$\frac{\partial \psi}{\partial \theta_i} = \eta_i$$

then satisfies

$$\mathrm{Hess}_\theta\, \psi = g.$$

Such a Riemannian manifold (M, g) is a Hessian manifold with potential ψ.

Let $\mathcal{N}$ denote the manifold of n-variate normal distributions

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right),$$

where

- $\mu \in \mathbb{R}^n$ is the mean,
- $\Sigma \in \mathrm{Pos}(n, \mathbb{R})$ is the covariance matrix.

Let $\mathcal{N}$ denote the manifold of n-variate normal distributions

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right),$$

where

- $\mu \in \mathbb{R}^n$ is the mean,
- $\Sigma \in \mathrm{Pos}(n, \mathbb{R})$ is the covariance matrix.

We choose coordinates $\theta = (\theta_i)$, $\Theta = (\Theta_{ij})$ on $\mathcal{N}$,

$$\Theta_{ij} = \Sigma_{ij}, \quad \theta_i = (\Sigma\mu)_i, \quad i, j = 1, \ldots, n.$$

The space $\mathcal{N}$ of normal distributions (I)

Let $\mathcal{N}$ denote the manifold of n-variate normal distributions

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right),$$

where

- $\mu \in \mathbb{R}^n$ is the mean,
- $\Sigma \in \mathrm{Pos}(n, \mathbb{R})$ is the covariance matrix.

We choose coordinates $\theta = (\theta_i)$, $\Theta = (\Theta_{ij})$ on $\mathcal{N}$,

$$\Theta_{ij} = \Sigma_{ij}, \quad \theta_i = (\Sigma\mu)_i, \quad i, j = 1, \dots, n.$$

Define

$$\psi(\theta, \Theta) = \frac{1}{2}(\theta^\top \Theta\theta - \log\det\Theta).$$

Let $\mathcal{N}$ denote the manifold of n-variate normal distributions

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right),$$

where

- $\mu \in \mathbb{R}^n$ is the mean,
- $\Sigma \in \mathrm{Pos}(n, \mathbb{R})$ is the covariance matrix.

We choose coordinates $\theta = (\theta_i)$, $\Theta = (\Theta_{ij})$ on $\mathcal{N}$,

$$\Theta_{ij} = \Sigma_{ij}, \quad \theta_i = (\Sigma\mu)_i, \quad i, j = 1, \ldots, n.$$

Define

$$\psi(\theta, \Theta) = \frac{1}{2}(\theta^\top \Theta\theta - \log\det\Theta).$$

Theorem
$\mathcal{N}$ becomes a statistical manifold with Fisher metric $g = \mathrm{Hess}_{\theta,\Theta}\,\psi$*,*
and $\mathcal{N}$ is ∇^{I}-flat.

The space $\mathcal{N}$ of normal distributions (II)

The geometry of $\mathcal{N}$:

- With the flat connection ∇^{I}, we identify $\mathcal{N}$ with the open convex cone $\mathbb{R}^{n} \times \mathrm{Pos}(n, \mathbb{R})$ in the vector space $\mathbb{R}^{n} \times \mathrm{Sym}(n, \mathbb{R})$ ($\cong \mathbb{R}^{n+\frac{n(n+1)}{2}}$).

The space $\mathcal{N}$ of normal distributions (II)

The geometry of $\mathcal{N}$:

- With the flat connection ∇^{I}, we identify $\mathcal{N}$ with the open convex cone $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$ in the vector space $\mathbb{R}^n \times \mathrm{Sym}(n, \mathbb{R})$ ($\cong \mathbb{R}^{n+\frac{n(n+1)}{2}}$).
- $\mathcal{N}$ splits further into a product of differentiable manifolds

$$\mathbb{R}^n \times \mathbb{R} \times \mathcal{P},$$

where $\mathcal{P} = \{\Sigma \in \mathrm{Pos}(n, \mathbb{R}) \mid \det \Sigma = 1\}$ and $\mathrm{Pos}(n, \mathbb{R}) = \mathbb{R} \times \mathcal{P}$.

The space $\mathcal{N}$ of normal distributions (II)

The geometry of $\mathcal{N}$:

- With the flat connection ∇^{I}, we identify $\mathcal{N}$ with the open convex cone $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$ in the vector space $\mathbb{R}^n \times \mathrm{Sym}(n, \mathbb{R})$ ($\cong \mathbb{R}^{n+\frac{n(n+1)}{2}}$).
- $\mathcal{N}$ splits further into a product of differentiable manifolds

$$\mathbb{R}^n \times \mathbb{R} \times \mathcal{P},$$

  where $\mathcal{P} = \{\Sigma \in \mathrm{Pos}(n, \mathbb{R}) \mid \det \Sigma = 1\}$ and $\mathrm{Pos}(n, \mathbb{R}) = \mathbb{R} \times \mathcal{P}$.
- Every $\Sigma \in \mathcal{P}$ can be written as $\Sigma = A^\top A$ for $A \in \mathrm{SL}(n, \mathbb{R})$.
  This means the Lie group $\mathrm{SL}(n, \mathbb{R})$ acts transitively on $\mathcal{P}$ by $A.\Sigma = A^\top \Sigma A$.

The space $\mathcal{N}$ of normal distributions (II)

The geometry of $\mathcal{N}$:

- With the flat connection ∇^{I}, we identify $\mathcal{N}$ with the open convex cone $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$ in the vector space $\mathbb{R}^n \times \mathrm{Sym}(n, \mathbb{R})$ ($\cong \mathbb{R}^{n+\frac{n(n+1)}{2}}$).
- $\mathcal{N}$ splits further into a product of differentiable manifolds

$$\mathbb{R}^n \times \mathbb{R} \times \mathcal{P},$$

  where $\mathcal{P} = \{\Sigma \in \mathrm{Pos}(n, \mathbb{R}) \mid \det \Sigma = 1\}$ and $\mathrm{Pos}(n, \mathbb{R}) = \mathbb{R} \times \mathcal{P}$.
- Every $\Sigma \in \mathcal{P}$ can be written as $\Sigma = A^\top A$ for $A \in \mathrm{SL}(n, \mathbb{R})$.
  This means the Lie group $\mathrm{SL}(n, \mathbb{R})$ acts transitively on $\mathcal{P}$ by $A.\Sigma = A^\top \Sigma A$.
- The stabilizer subgroup of this action at $\Sigma = I_n$ is $\mathrm{SO}(n)$.
  Hence we identify $\mathcal{P}$ with the homogeneous space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.

The space $\mathcal{N}$ of normal distributions (II)

The geometry of $\mathcal{N}$:

- With the flat connection ∇^{I}, we identify $\mathcal{N}$ with the open convex cone $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$ in the vector space $\mathbb{R}^n \times \mathrm{Sym}(n, \mathbb{R})$ ($\cong \mathbb{R}^{n+\frac{n(n+1)}{2}}$).
- $\mathcal{N}$ splits further into a product of differentiable manifolds

$$\mathbb{R}^n \times \mathbb{R} \times \mathcal{P},$$

  where $\mathcal{P} = \{\Sigma \in \mathrm{Pos}(n, \mathbb{R}) \mid \det \Sigma = 1\}$ and $\mathrm{Pos}(n, \mathbb{R}) = \mathbb{R} \times \mathcal{P}$.
- Every $\Sigma \in \mathcal{P}$ can be written as $\Sigma = A^\top A$ for $A \in \mathrm{SL}(n, \mathbb{R})$.
  This means the Lie group $\mathrm{SL}(n, \mathbb{R})$ acts transitively on $\mathcal{P}$ by $A.\Sigma = A^\top \Sigma A$.
- The stabilizer subgroup of this action at $\Sigma = I_n$ is $\mathrm{SO}(n)$.
  Hence we identify $\mathcal{P}$ with the homogeneous space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.
- The Fisher metric $\mathrm{g}^{\mathcal{P}}$ restricted to $\mathcal{P}$ equals

$$\mathrm{g}^{\mathcal{P}}_{\Sigma}(X, Y) = \mathrm{tr}(\Sigma^{-1} X \Sigma^{-1} Y).$$

  This means $(\mathcal{P}, \mathrm{g}^{\mathcal{P}})$ is the Riemannian symmetric space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$ with metric induced by the Killing form of $\mathrm{SL}(n, \mathbb{R})$.

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

- This is equivalent to the existence of a point reflection of geodesic curves at each point $p \in M$.

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

- This is equivalent to the existence of a point reflection of geodesic curves at each point $p \in M$.
- Riemannian symmetric spaces were fully classified by Élie Cartan in 1926.

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

- This is equivalent to the existence of a point reflection of geodesic curves at each point $p \in M$.
- Riemannian symmetric spaces were fully classified by Élie Cartan in 1926.
- Every (non-Euclidean) Riemannian symmetric space is a Riemannian product of irreducible Riemannian symmetric spaces, which are either
  - a simple Lie group G, or
  - a quotient G/K of a simple Lie group by a maximal compact subgroup K (e.g. $G = \mathrm{SL}(n, \mathbb{R})$ and $K = \mathrm{SO}(n)$).

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

- This is equivalent to the existence of a point reflection of geodesic curves at each point $p \in M$.
- Riemannian symmetric spaces were fully classified by Élie Cartan in 1926.
- Every (non-Euclidean) Riemannian symmetric space is a Riemannian product of irreducible Riemannian symmetric spaces, which are either
  - a simple Lie group G, or
  - a quotient G/K of a simple Lie group by a maximal compact subgroup K (e.g. $G = \mathrm{SL}(n, \mathbb{R})$ and $K = \mathrm{SO}(n)$).
- The metric on the symmetric space comes from a bi-invariant metric on G (meaning left- and right-multiplication on G are isometries).

A Riemannian symmetric space is a simply connected Riemannian manifold M such that $\nabla \mathrm{R} = 0$ (lax: M looks the same everywhere).

- This is equivalent to the existence of a point reflection of geodesic curves at each point $p \in M$.
- Riemannian symmetric spaces were fully classified by Élie Cartan in 1926.
- Every (non-Euclidean) Riemannian symmetric space is a Riemannian product of irreducible Riemannian symmetric spaces, which are either
  - a simple Lie group G, or
  - a quotient G/K of a simple Lie group by a maximal compact subgroup K (e.g. $G = \mathrm{SL}(n, \mathbb{R})$ and $K = \mathrm{SO}(n)$).
- The metric on the symmetric space comes from a bi-invariant metric on G (meaning left- and right-multiplication on G are isometries).

Now back to $\mathcal{N}$...

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$.
- $\mathbb{R}$ and $\mathbb{R}^n$ with their canonical scalar products are symmetric spaces of Euclidean type.

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.
- $\mathbb{R}$ and $\mathbb{R}^n$ with their canonical scalar products are symmetric spaces of Euclidean type.
- $\mathrm{Pos}(n, \mathbb{R})$ with the restricted Fisher metric is also a symmetric space

$$\mathrm{GL}(n, \mathbb{R})/\mathrm{O}(n) = \mathbb{R} \times \mathcal{P},$$

  the Riemannian product of $\mathbb{R}$ and $\mathcal{P}$.

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.
- $\mathbb{R}$ and $\mathbb{R}^n$ with their canonical scalar products are symmetric spaces of Euclidean type.
- $\mathrm{Pos}(n, \mathbb{R})$ with the restricted Fisher metric is also a symmetric space

  $$\mathrm{GL}(n, \mathbb{R})/\mathrm{O}(n) = \mathbb{R} \times \mathcal{P},$$

  the Riemannian product of $\mathbb{R}$ and $\mathcal{P}$.
- However, the restriction of g to $\mathbb{R}^n \cong \{(\mu, \Sigma) \mid \mu \in \mathbb{R}^n\}$ depends on Σ. Hence $\mathcal{N}$ is not a Riemannian product $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$.

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$.
- $\mathbb{R}$ and $\mathbb{R}^n$ with their canonical scalar products are symmetric spaces of Euclidean type.
- $\mathrm{Pos}(n,\mathbb{R})$ with the restricted Fisher metric is also a symmetric space

  $$\mathrm{GL}(n,\mathbb{R})/\mathrm{O}(n) = \mathbb{R} \times \mathcal{P},$$

  the Riemannian product of $\mathbb{R}$ and $\mathcal{P}$.
- However, the restriction of g to $\mathbb{R}^n \cong \{(\mu,\Sigma) \mid \mu \in \mathbb{R}^n\}$ depends on Σ. Hence $\mathcal{N}$ is not a Riemannian product $\mathbb{R}^n \times \mathrm{Pos}(n,\mathbb{R})$.

Theorem

$\mathcal{N}$ is a trivial vector bundle

$$\mathbb{R}^n \longrightarrow \mathcal{N} \longrightarrow \mathrm{Pos}(n,\mathbb{R})$$

where fiber $\mathbb{R}^n$ and base $\mathrm{Pos}(n,\mathbb{R})$ are symmetric spaces.

The space $\mathcal{N}$ of normal distributions (III)

More geometry of $\mathcal{N}$:

- $\mathcal{P}$ is the symmetric space $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$.
- $\mathbb{R}$ and $\mathbb{R}^n$ with their canonical scalar products are symmetric spaces of Euclidean type.
- $\mathrm{Pos}(n, \mathbb{R})$ with the restricted Fisher metric is also a symmetric space

  $$\mathrm{GL}(n, \mathbb{R})/\mathrm{O}(n) = \mathbb{R} \times \mathcal{P},$$

  the Riemannian product of $\mathbb{R}$ and $\mathcal{P}$.
- However, the restriction of g to $\mathbb{R}^n \cong \{(\mu, \Sigma) \mid \mu \in \mathbb{R}^n\}$ depends on Σ. Hence $\mathcal{N}$ is not a Riemannian product $\mathbb{R}^n \times \mathrm{Pos}(n, \mathbb{R})$.

Theorem

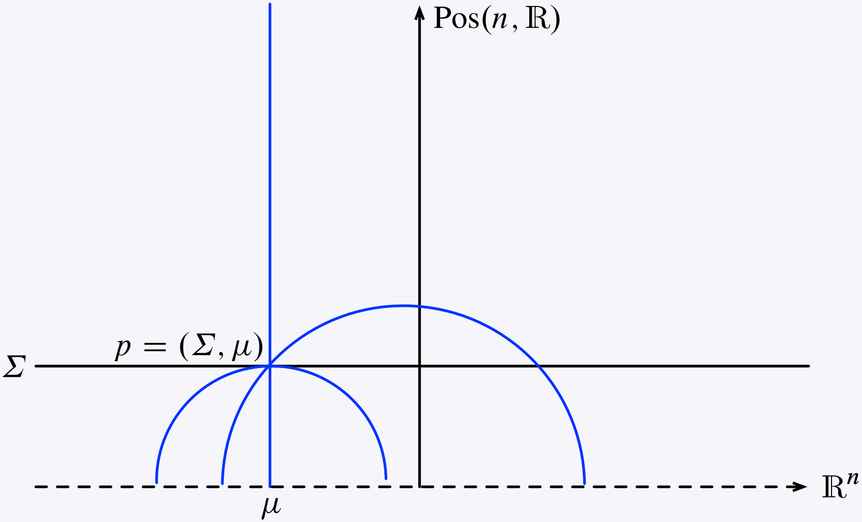
$\mathcal{N}$ is a trivial vector bundle

$$\mathbb{R}^n \longrightarrow \mathcal{N} \longrightarrow \mathrm{Pos}(n, \mathbb{R})$$

where fiber $\mathbb{R}^n$ and base $\mathrm{Pos}(n, \mathbb{R})$ are symmetric spaces.

For $n = 1$, $\mathcal{N}$ with the Fisher metric equals the hyperbolic plane.

The space $\mathcal{N}$ of normal distributions (IV)



Outlook

- Many other classes of statistical manifolds (exponential families, distributions on finite sets,...) have similar properties (flatness, actions by Lie groups,...).

Outlook

- Many other classes of statistical manifolds (exponential families, distributions on finite sets,. . .) have similar properties (flatness, actions by Lie groups,. . .).
- Hessian manifolds are a real analogue of Kähler manifolds.
  Quantum information theory uses Kähler metrics.

Outlook

- Many other classes of statistical manifolds (exponential families, distributions on finite sets,...) have similar properties (flatness, actions by Lie groups,...).
- Hessian manifolds are a real analogue of Kähler manifolds.
  Quantum information theory uses Kähler metrics.
- There is a well-developed theory of Hessian manifolds.
  Converse question: Which Hessian manifolds are statistical manifolds?

- Many other classes of statistical manifolds (exponential families, distributions on finite sets,...) have similar properties (flatness, actions by Lie groups,...).
- Hessian manifolds are a real analogue of Kähler manifolds.
  Quantum information theory uses Kähler metrics.
- There is a well-developed theory of Hessian manifolds.
  Converse question: Which Hessian manifolds are statistical manifolds?
- "Pseudo-statistics": Homogeneous space with indefinite Hessian metrics.
  Does a non-positive definite Fisher metric make any sense from a statistical point of view?