Mini-Course on Information Geometry

Introduction

Herlock Rahimi
Department of Electrical and Computer Engineering
Yale University

June 5, 2025

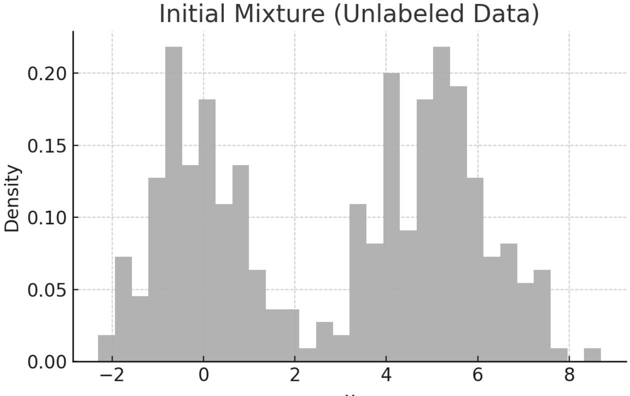
Overview

1. GMM: Gaussian Mixture Model
2. Riemannian Geometry
3. Exponentiation Family and Sufficient Statistics
4. Riemannian Geometry as an extension of Straight line
5. Autoparallel Transport
6. Geodesics
7. Curvature
8. Connection and Riemannian Metric
9. Bregman Divergence
10. Duallity
11. Mirror Descent
12. In Reinforcement Learning

Two-Gaussian Mixture Model: Problem Setup

Goal: Given data from two unlabelled Gaussian distributions, estimate parameters $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$.

- Data: $x_1, \ldots, x_n \sim \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2)$
- Challenge: Data is *unlabeled*; we don't know which point came from which Gaussian.
- Direct MLE is intractable $\Rightarrow$ use EM algorithm.



EM Algorithm: Soft Assignment (Mathematical Steps)

Problem: Maximize log-likelihood of a mixture of Gaussians:

$$\log L(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2) \right)$$

E-Step: Compute **responsibilities** (soft cluster assignments):

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)}$$

M-Step: Maximize expected complete-data log-likelihood:

$$\pi_k^{\text{new}} = \frac{1}{n} \sum_{i=1}^{n} r_{ik}, \quad \mu_k^{\text{new}} = \frac{\sum_{i=1}^{n} r_{ik} x_i}{\sum_{i=1}^{n} r_{ik}}, \quad \sigma_k^2 = \frac{\sum_{i=1}^{n} r_{ik}(x_i - \mu_k)^2}{\sum_{i=1}^{n} r_{ik}}$$

Repeat until convergence.

Non-EM Algorithm: Hard Assignment (Heuristic MLE)

Heuristic: Alternate between hard clustering and parameter updates.
Initialize: Random hard assignments $z_i \in \{0, 1\}$.
M-Step: Update parameters using current assignments:

$$\pi_k = \frac{n_k}{n}, \quad \mu_k = \frac{1}{n_k} \sum_{i:z_i=k} x_i, \quad \sigma_k^2 = \frac{1}{n_k} \sum_{i:z_i=k} (x_i - \mu_k)^2$$

E-Step: Reassign each point to the most likely Gaussian:

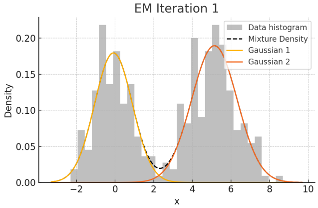
$$z_i = \arg\max_k \left[\pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)\right]$$

Repeat for fixed number of iterations or until assignments stop changing.

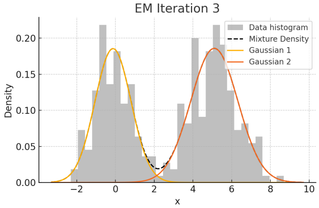
Illustration of EM Iterations

Initial guess: Random parameters.

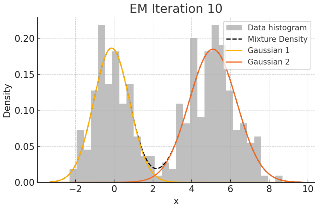
Iteration 1:



Iteration 3:



Iteration 10 (Converged):

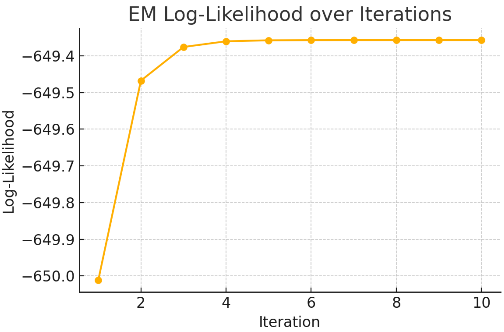


Each iteration updates the soft assignment and shifts the Gaussian parameters closer to the true generating process.

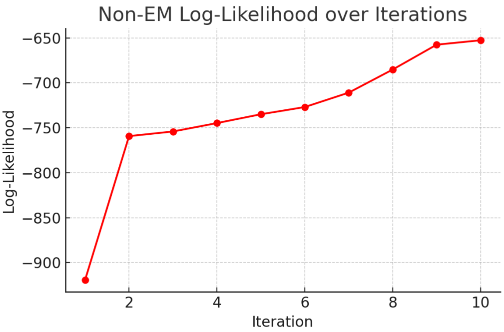
Comparison: EM vs. Non-EM Learning

Experiment: Run EM and naive hard assignment (random or k-means-like) for 10 iterations.

EM Algorithm



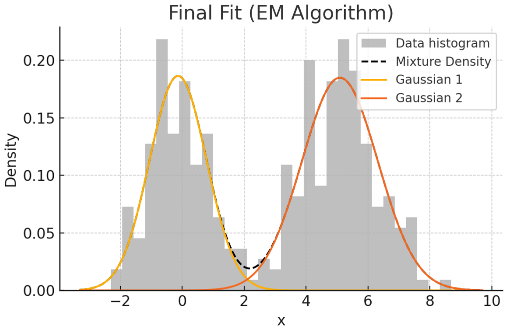
Non-EM Assignment



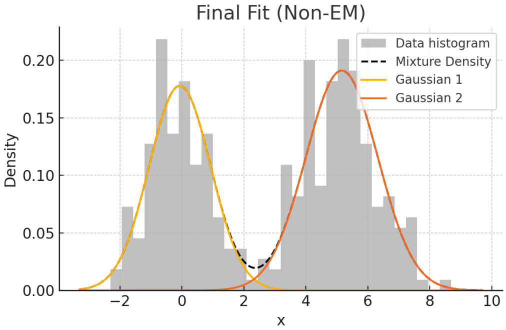
Observation: EM converges more quickly and smoothly; non-EM can oscillate or diverge.

Learned Distributions: EM vs. Non-EM

EM Learned Distributions



Non-EM Learned Distributions



EM fit is closer to the true generating distributions.

Smooth Manifolds

Definition

A **smooth manifold** $\mathcal{M}$ of dimension n is a topological space that is:

- Hausdorff and second-countable
- Locally homeomorphic to $\mathbb{R}^n$
- Equipped with a maximal smooth atlas

Charts: Each point $p \in \mathcal{M}$ has a neighborhood $U \subset \mathcal{M}$ and a homeomorphism (chart)

$$\varphi : U \to \varphi(U) \subset \mathbb{R}^n$$

such that transition maps $\varphi_j \circ \varphi_i^{-1}$ are C^∞ smooth where defined.

Tangent Vectors

Definition

Let $\mathcal{M}$ be a smooth manifold. A **tangent vector** at a point $p \in \mathcal{M}$ is a derivation:

$$v : C^\infty(\mathcal{M}) \to \mathbb{R}$$

such that:

$$v(fg) = v(f)g(p) + f(p)v(g)$$

for all $f, g \in C^\infty(\mathcal{M})$. This is the Leibniz rule.

Notation: The set of all tangent vectors at p forms a real vector space:

$$T_p\mathcal{M} := \text{Tangent space at } p$$

Tangent Space in Coordinates

Given a chart (U, φ), where $\varphi(p) = (x^1, \ldots, x^n) \in \mathbb{R}^n$, the basis of $T_p\mathcal{M}$ is:

$$\left\{ \frac{\partial}{\partial x^i}\bigg|_p \right\}_{i=1}^{n}$$

defined via:

$$\frac{\partial}{\partial x^i}\bigg|_p (f) := \frac{\partial (f \circ \varphi^{-1})}{\partial x^i}\bigg|_{\varphi(p)}$$

Any tangent vector can be written:

$$v = v^i \frac{\partial}{\partial x^i}\bigg|_p$$

for some $v^i \in \mathbb{R}$.

Vector Fields

Definition

A **smooth vector field** on $\mathcal{M}$ assigns to each point $p \in \mathcal{M}$ a tangent vector $X_p \in T_p\mathcal{M}$, smoothly varying with p.

Space of vector fields:

$$\mathfrak{X}(\mathcal{M}) := \text{set of smooth sections } X : \mathcal{M} \to T\mathcal{M}$$

In coordinates x^i, a vector field has the local form:

$$X = X^i(x)\frac{\partial}{\partial x^i}$$

where $X^i \in C^\infty(\mathcal{M})$.

Riemannian Metric

Definition

A **Riemannian metric** on a smooth manifold $\mathcal{M}$ is a smooth assignment:

$$g : p \mapsto g_p$$

where each g_p is a positive-definite inner product on $T_p\mathcal{M}$, such that:

$$g_p(v, w) = g_q(\phi_* v, \phi_* w)$$

under smooth coordinate changes.

In coordinates:

$$g = g_{ij}(x)\, dx^i \otimes dx^j \quad \text{where} \quad g_{ij}(x) = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$$

and $g_{ij}(x)$ is symmetric and positive-definite for all x.

Norms and Inner Products

Let $v \in T_p\mathcal{M}$. The Riemannian metric g induces:

- An inner product:

$$\langle v, w \rangle_p := g_p(v, w)$$

- A norm:

$$\|v\| := \sqrt{g_p(v, v)}$$

- The angle between vectors:

$$\cos\theta = \frac{g_p(v, w)}{\|v\|\|w\|}$$

Hence, a Riemannian metric generalizes Euclidean geometry to smooth manifolds.

Pullback Metric and Charts

Let (U, φ) be a chart with coordinates x^i, and let φ_* be the pushforward. Then the metric in local coordinates becomes:

$$g_{ij}(x) = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$$

If $f : \mathcal{M} \to \mathbb{R}$ is smooth, then the gradient is:

$$\mathrm{grad}_g f = g^{ij} \frac{\partial f}{\partial x^j} \frac{\partial}{\partial x^i}$$

where g^{ij} is the inverse matrix of g_{ij}.

Summary and Road Ahead

- A **smooth manifold** provides a coordinate-free generalization of Euclidean space.
- The **tangent space** $T_p\mathcal{M}$ is the linearization of $\mathcal{M}$ at a point.
- A **Riemannian metric** equips each tangent space with an inner product.
- These structures enable us to define geometry on abstract manifolds: angles, lengths, gradients, and more.

Next: Connections, covariant derivatives, geodesics, curvature, and how they connect to statistical models.

Exponential Family: Definition and Structure (I/II)

Definition: A family of probability distributions is an **exponential family** if it can be written as:

$$p(x;\theta) = h(x) \exp\left(\sum_{i=1}^{d} \theta^i F_i(x) - \psi(\theta)\right)$$

where:

- $\theta = (\theta^1, \ldots, \theta^d)$: **natural parameters** (e-coordinates)
- $F(x) = (F_1(x), \ldots, F_d(x))$: **sufficient statistics**
- $\psi(\theta)$: **log-partition function**
- $h(x)$: base measure

Dual coordinates:

$$\eta_i := \mathbb{E}_\theta[F_i(x)] \quad \text{(m-coordinates)}$$

Duality: $\eta = \nabla\psi(\theta)$, and $\theta = \nabla\varphi(\eta)$, where φ is the Legendre dual of ψ.

Examples of Exponential Families (II/II)

1. **Bernoulli(θ)**

$$p(x;\theta) = \theta^x(1-\theta)^{1-x} = \exp\left(x \log \frac{\theta}{1-\theta} + \log(1-\theta)\right)$$

Sufficient statistic: $F(x) = x$, $\quad$ e-param: $\theta^{(e)} = \log \frac{\theta}{1-\theta}$

2. **Gaussian(μ, σ^2)**

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

Exponential form:

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2}, \quad F(x) = (x, x^2)$$

3. **Poisson(λ)**

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp\left(x \log\lambda - \lambda - \log x!\right)$$

Sufficient statistic: $F(x) = x$, $\quad \theta = \log\lambda$, $\quad \eta = \lambda$

Motivation: Why Legendre Transform and Bregman Divergence? (I/IV)

We want:

- A principled way to measure "distance" without requiring symmetry
- Geometry adapted to convex optimization and information
- Tools for dual coordinate systems (e.g., exponential vs. expectation)

Key tools:

- **Convex functions** define natural dual spaces
- **Legendre transform** maps between these dual spaces
- **Bregman divergence** measures mismatch based on convexity

[Insert figure: Convex function with tangent at a point and dual axis]

Motivation for Connections (I/II)

Problem: In Riemannian geometry, we know how to measure lengths and angles, but how do we compare vectors at different points on a manifold?

Analogy: Imagine walking on Earth's surface holding an arrow. As you walk along a path, you keep the arrow "pointing in the same direction" — but what does that mean on a curved surface?

Need: We need a way to define how vectors change along curves — this leads to the concept of a **connection**.

In Information Geometry: Consider the statistical manifold of 1D normal distributions. How do we move a tangent vector (infinitesimal change in parameters) from one distribution to another while preserving its "meaning"?

Motivation for Connections (II/II)

Statistical Manifold: The set of 1D normal distributions $\mathcal{N}(\mu, \sigma^2)$ forms a 2D differentiable manifold.

Two coordinate systems:

- **Exponential (e-) coordinates:** $(\theta^1 = \mu/\sigma^2, \theta^2 = -1/(2\sigma^2))$
- **Mixture (m-) coordinates:** $(\eta^1 = \mu, \eta^2 = \mu^2 + \sigma^2)$

Two kinds of "straight lines":

- **e-connection ($\alpha = 1$):** Geodesics correspond to exponential families.
- **m-connection ($\alpha = -1$):** Geodesics correspond to mixture families.

Affine Connection: Formal Definition (I/II)

Definition: An **affine connection** ∇ on a differentiable manifold M assigns to each pair of vector fields X, Y a new vector field $\nabla_X Y$ satisfying:

1. Linearity in X: $\nabla_{fX+gZ} Y = f\nabla_X Y + g\nabla_Z Y$
2. Leibniz in Y: $\nabla_X(fY) = (Xf)Y + f\nabla_X Y$
3. Linearity in Y

Interpretation: $\nabla_X Y$ tells us how the vector field Y changes in the direction of X — this is the infinitesimal version of parallel transport.

Connections in Coordinates (II/II)

Let $(x^1, \ldots, x^n)$ be a coordinate system on M. The connection is determined by the **Christoffel symbols** Γ_{ij}^k via:

$$\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}^k \partial_k$$

For any vector fields $X = X^i\partial_i$, $Y = Y^j\partial_j$, we get:

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

In Information Geometry: We define a family of α-connections where the Christoffel symbols depend on the underlying statistical structure (Fisher metric, etc.).

Example: Exponential Connection ($\alpha = 1$)

Statistical Model: 1D normal distribution $\mathcal{N}(\mu, \sigma^2)$

Exponential coordinates:

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2}$$

Fisher Metric:

$$g_{ij}(\theta) = \mathbb{E}_\theta \left[\partial_i \log p(x;\theta) \partial_j \log p(x;\theta)\right]$$

Christoffel symbols (e-connection):

$$\Gamma^{(e)}_{ijk} = \mathbb{E}\left[\partial_i \partial_j \log p(x;\theta) \partial_k \log p(x;\theta)\right]$$

Example: Mixture Connection ($\alpha = -1$)

Mixture coordinates:

$$\eta^1 = \mu, \quad \eta^2 = \mu^2 + \sigma^2$$

Christoffel symbols (m-connection):

$$\Gamma^{(m)}_{ijk} = -\mathbb{E}\left[\partial_k \partial_i \log p(x;\theta) \partial_j \log p(x;\theta)\right]$$

Interpretation: The m-geodesics correspond to linear combinations of distributions (mixtures), e.g., convex combinations of Gaussians.

Motivation for Autoparallel Transport (I/II)

Core Question: How do we "move" a vector from one point on a manifold to another in a way that preserves its direction relative to the manifold's geometry?

Flat space intuition: In $\mathbb{R}^n$, we can translate a vector unchanged. On curved manifolds, translation is not intrinsic — we need a rule for consistent movement: **autoparallel transport**.

Given: A smooth curve $\gamma(t)$ on a manifold M and a vector $V(t)$ "attached" at each point $\gamma(t)$.

Goal: Describe how to evolve $V(t)$ along $\gamma(t)$ so that it stays "parallel" with respect to the connection.

Motivation for Autoparallel Transport (II/II)

Transporting a vector field $V(t)$ along a path $\gamma(t)$:

- We use the connection ∇ to define the derivative of $V(t)$ along $\gamma(t)$.
- Autoparallel transport demands that this derivative vanishes:

$$\nabla_{\dot{\gamma}(t)} V(t) = 0$$

Interpretation: $V(t)$ doesn't "twist" or "rotate" with respect to the geometry induced by ∇.

Information Geometry Viewpoint: This process lets us understand how local changes in parameters (e.g., score functions) evolve under statistical flows.

Definition: Autoparallel Transport (I/II)

Let $\gamma : I \to M$ be a smooth curve on manifold M, and let ∇ be an affine connection.

A vector field $V(t)$ along $\gamma(t)$ is said to be **autoparallel transported** (or just **parallel transported**) if:

$$\nabla_{\dot{\gamma}(t)} V(t) = 0$$

Interpretation: The vector $V(t)$ maintains a constant direction relative to the geometry defined by ∇.

Initial condition: Given $V(t_0) = V_0$, the equation has a unique solution — i.e., transport is well-defined.

Autoparallel Transport in Coordinates (II/II)

In local coordinates: Let $V(t) = V^k(t)\partial_k$ be a vector field along $\gamma(t) = (x^i(t))$.

Then the autoparallel condition becomes:

$$\frac{dV^k}{dt} + \sum_{i,j} \Gamma_{ij}^k(x(t))\frac{dx^i}{dt}V^j(t) = 0$$

Linear ODE system: This is a first-order linear differential equation system for $V^k(t)$ with smooth coefficients.

In Riemannian geometry: Parallel transport preserves the inner product; in general α-connections, it need not.

Example: Autoparallel Transport with e-Connection ($\alpha = 1$)

Model: Gaussian $\mathcal{N}(\mu, \sigma^2)$, exponential coordinates:

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2}$$

Let $\gamma(t)$ be an exponential geodesic between two distributions. Consider a vector field $V(t)$ along $\gamma(t)$ representing change in sufficient statistics.

Autoparallel transport with respect to $\nabla^{(1)}$ satisfies:

$$\frac{dV^k}{dt} + \Gamma_{ij}^{(1)k} \frac{d\theta^i}{dt} V^j = 0$$

This describes how natural parameters evolve in an exponential family under constant statistical "direction".

Example: Autoparallel Transport with m-Connection ($\alpha = -1$)

Mixture coordinates:

$$\eta^1 = \mu, \quad \eta^2 = \mu^2 + \sigma^2$$

Let $\gamma(t)$ be a path representing a convex combination of Gaussians.

Transport a score function or Fisher information direction $V(t)$ along $\gamma(t)$ with respect to $\nabla^{(-1)}$:

$$\frac{dV^k}{dt} + \Gamma_{ij}^{(-1)k} \frac{d\eta^i}{dt} V^j = 0$$

This models how expectation parameters change under blending distributions — crucial in EM-type algorithms.

Motivation for Geodesics (I/II)

Basic idea: In Euclidean space, the shortest path between two points is a straight line.

On a curved manifold: The shortest path generalizes to a **geodesic** — a curve that "locally minimizes distance".

But in Information Geometry, we also care about curves that look straight in coordinate systems defined by statistical structure.

Key Question: What kind of "straightness" do we want — metric or affine?

Motivation for Geodesics (II/II)

Two views on geodesics:

- **Metric view (Levi-Civita):** Geodesics locally minimize distance w.r.t. Fisher metric.
- **Affine view (Connection-based):** Geodesics are **autoparallel** curves under a given connection.

In Information Geometry:

- e-geodesics: straight in exponential coordinates
- m-geodesics: straight in mixture coordinates

Geometrical insight: These geodesics reflect natural statistical paths (MLE paths, convex mixtures, etc.)

Let $\gamma : I \to M$ be a smooth curve on a manifold M with affine connection ∇.

Then γ is a **geodesic** (w.r.t. ∇) if:

$$\nabla_{\dot{\gamma}(t)} \dot{\gamma}(t) = 0$$

This is equivalent to: the velocity vector is autoparallel transported along the curve.

Interpretation: No external "acceleration" — the motion is natural to the geometry.

Geodesics in Local Coordinates (II/II)

Let $\gamma(t) = (x^1(t), \ldots, x^n(t))$ in local coordinates. Then the geodesic equation becomes:

$$\frac{d^2x^k}{dt^2} + \sum_{i,j} \Gamma^k_{ij}(x(t)) \frac{dx^i}{dt} \frac{dx^j}{dt} = 0$$

This is a second-order nonlinear ODE system determined by the connection ∇.

In Riemannian geometry: Use the Levi-Civita connection from the Fisher metric.

In Info Geometry: Use α-connections for $\alpha \in [-1, 1]$.

Example: e-Geodesics ($\alpha = 1$)

Model: $\mathcal{N}(\mu, \sigma^2)$ with exponential coordinates:

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2}$$

e-geodesic:

$$\theta(t) = (1 - t)\theta_0 + t\theta_1$$

A straight line in exponential coordinates.

Interpretation: This corresponds to paths within exponential families — natural for likelihood-based inference.

Transport: Velocity vector remains parallel under $\nabla^{(1)}$.

Example: m-Geodesics ($\alpha = -1$)

Mixture coordinates:

$$\eta^1 = \mu, \quad \eta^2 = \mu^2 + \sigma^2$$

m-geodesic:

$$\eta(t) = (1 - t)\eta_0 + t\eta_1$$

A straight line in mixture coordinates.

Interpretation: This represents convex combinations of probability distributions.

Transport: Velocity remains parallel under $\nabla^{(-1)}$.

Motivation for Curvature (I/II)

Intuition: Flat space allows us to move vectors around without distortion. Curved space does not.

Thought experiment: Move a vector around a loop on a sphere — it comes back rotated. Something intrinsic to the space caused this change.

Key Question: How does a manifold "resist" the parallel transport of vectors?

Answer: This failure to return the same vector defines **curvature**, and we express it using the connection ∇.

Visual intuition:

- In flat space: vector transported around a loop remains unchanged.
- In curved space: the direction changes — something geometric caused a "twist".

Why it matters in Information Geometry:

- The statistical manifold may be flat under one connection (e.g., $\nabla^{(1)}$) and curved under another (e.g., Levi-Civita).
- Curvature determines whether geodesics can intersect or deviate.

Definition: Curvature Tensor (I/II)

Given: An affine connection ∇ on a manifold M.

The **curvature tensor** R is defined by:

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

Interpretation:

- Measures the failure of second covariant derivatives to commute.
- Describes the infinitesimal effect of parallel transporting Z around a parallelogram spanned by X, Y.

In coordinates: The curvature tensor has components:

$$R^l{}_{ijk} = \partial_j \Gamma^l_{ik} - \partial_i \Gamma^l_{jk} + \Gamma^m_{ik}\Gamma^l_{jm} - \Gamma^m_{jk}\Gamma^l_{im}$$

Symmetries:

$$R(X, Y) = -R(Y, X)$$

Flatness: A connection ∇ is **flat** if $R = 0$ everywhere.

In Info Geometry:

- Both $\nabla^{(1)}$ and $\nabla^{(-1)}$ are flat.
- The Levi-Civita connection (from Fisher metric) is curved.

Example: Flatness of e-Connection ($\alpha = 1$)

Statistical manifold: 1D Gaussians, exponential coordinates:

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2}$$

e-Connection:

$$\Gamma^{(1)}_{ijk} = \mathbb{E}\left[\partial_i \partial_j \log p(x;\theta) \partial_k \log p(x;\theta)\right]$$

Result: $R^{(1)} = 0$ — the manifold is flat under $\nabla^{(1)}$.

Implication: Exponential coordinate system behaves like affine space — geodesics are straight lines.

Motivation: How Does a Connection Relate to a Metric? (I/II)

So far:

- The **Riemannian metric** (Fisher metric) lets us measure lengths, angles, and distances.
- A **connection** tells us how to compare or transport vectors across different points.

Question: Can the connection be consistent with the metric?

Example: In Riemannian geometry, the **Levi-Civita connection** is compatible with the metric:

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

But in Information Geometry: We often use non-metric connections (e.g., $\nabla^{(\alpha)}$) — they don't preserve the metric.

Motivation: Duality of Connections (II/II)

What if we allow for two connections, ∇ and ∇^*?

Then we can ask them to be "dual" with respect to the metric g:

$$X[g(Y, Z)] = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$$

Interpretation:

- ∇ distorts the metric in one direction.
- ∇^* undoes that distortion in the other direction.

In Information Geometry: The α-connections are **mutually dual**:

$$(\nabla^{(\alpha)})^* = \nabla^{(-\alpha)}$$

Definition: Metric Compatibility and Duality (I/II)

Metric compatibility: A connection ∇ is metric-compatible if:

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

Levi-Civita connection is the unique connection that is:

- **Torsion-free**
- **Metric-compatible**

Dual connections: A pair (∇, ∇^*) is **dual** w.r.t. a Riemannian metric g if:

$$X[g(Y,Z)] = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$$

Example: Duality on the Gaussian Manifold

Manifold: $\mathcal{N}(\mu, \sigma^2)$

Coordinate systems:

- **e-coordinates:** $\theta = (\mu/\sigma^2, -1/(2\sigma^2))$
- **m-coordinates:** $\eta = (\mu, \mu^2 + \sigma^2)$

Observation:

- $\nabla^{(1)}$: flat in θ-coordinates
- $\nabla^{(-1)}$: flat in η-coordinates
- $\nabla^{(0)}$: Levi-Civita (curved in both)

These two connections are dual under the Fisher metric.

Summary: Geometry of Dual Connections

Three central connections:

- $\nabla^{(1)}$: flat in exponential coordinates
- $\nabla^{(-1)}$: flat in mixture coordinates
- $\nabla^{(0)}$: Levi-Civita, metric-compatible

Duality:

$$\nabla^{(\alpha)} \text{ and } \nabla^{(-\alpha)} \text{ are dual w.r.t. the Fisher metric}$$

Implications:

- Dual geodesics intersect orthogonally under the Fisher metric.
- Projections (e.g., MLE vs. moment matching) follow dual geodesics.

Definitions: Legendre Transform and Bregman Divergence (II/IV)

Legendre Transform: For strictly convex $\psi : \mathbb{R}^n \to \mathbb{R}$, define:

$$\varphi(\eta) := \sup_{\theta} \{\langle \theta, \eta \rangle - \psi(\theta)\}$$

Then:

$$\eta = \nabla\psi(\theta), \quad \theta = \nabla\varphi(\eta)$$

Bregman Divergence:

$$D_\psi(\theta \| \theta') := \psi(\theta) - \psi(\theta') - \langle \nabla\psi(\theta'), \theta - \theta' \rangle$$

Asymmetry: In general, $D_\psi(\theta\|\theta') \neq D_\psi(\theta'\|\theta)$

[Insert figure: Gap between function and tangent plane at θ']

Duality of Bregman Divergences:

Given a convex function $\psi(\theta)$ with Legendre dual $\varphi(\eta)$:

$$D_\psi(\theta\|\theta') = \psi(\theta) + \varphi(\eta') - \langle \theta, \eta' \rangle$$

$$D_\varphi(\eta'\|\eta) = \varphi(\eta') + \psi(\theta) - \langle \eta', \theta \rangle$$

They are symmetric under duality:

$$D_\psi(\theta\|\theta') = D_\varphi(\eta'\|\eta)$$

Example: Negative Entropy

$$\psi(p) = \sum_i p_i \log p_i \quad \text{(negative entropy)}$$

$$D_\psi(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} = D_{\mathrm{KL}}(p\|q)$$

Legendre dual: $\varphi(u) = \log \sum_i \exp(u_i) \to$ softmax potential

Statistical Manifolds

Let $\mathcal{M} = \{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^n\}$ be a family of probability density functions with smooth dependence on the parameter θ.

Definition: Statistical Manifold

A statistical manifold is a differentiable manifold $\mathcal{M}$ where each point corresponds to a probability distribution $p(x;\theta)$, and the parameter space Θ serves as a coordinate chart.

We endow $\mathcal{M}$ with:

- A Riemannian metric g (Fisher information metric)
- Two affine connections $\nabla^{(e)}$, $\nabla^{(m)}$

Kullback–Leibler Divergence

Definition

The Kullback–Leibler divergence between two distributions $p(x;\theta)$ and $p(x;\theta')$ is defined as:

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta'}) = \int p(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta')}\, dx$$

- D_{KL} is not symmetric
- $D_{\mathrm{KL}}(p_\theta \| p_{\theta'}) \geq 0$, with equality iff $\theta = \theta'$
- Used to define both the Riemannian metric and affine connections on $\mathcal{M}$

Fisher Information Metric

The Fisher information matrix is defined by:

$$g_{ij}(\theta) := \mathbb{E}_\theta\left[\frac{\partial \log p(x;\theta)}{\partial \theta^i}\frac{\partial \log p(x;\theta)}{\partial \theta^j}\right]$$

Equivalently (from KL divergence):

$$g_{ij}(\theta) = \left.\frac{\partial^2}{\partial \theta^i \partial \theta'^j} D_{\mathrm{KL}}(p_\theta \| p_{\theta'})\right|_{\theta'=\theta}$$

This defines a Riemannian metric on $\mathcal{M}$, and is invariant under sufficient statistics (invariance principle).

Affine Connections from KL Divergence

KL divergence defines affine connections via third-order derivatives:

Christoffel Symbols of e-Connection ($\nabla^{(e)}$)

$$\Gamma^{(e)}_{ijk} := -\frac{\partial^3}{\partial\theta^i \partial\theta^j \partial\theta'^k} D_{\mathrm{KL}}(p_\theta \| p_{\theta'})\Big|_{\theta'=\theta}$$

$$\Gamma^{(e)k}_{ij} = g^{kl}\Gamma^{(e)}_{ijl}$$

Christoffel Symbols of m-Connection ($\nabla^{(m)}$)

$$\Gamma^{(m)}_{ijk} := -\frac{\partial^3}{\partial\theta'^i \partial\theta'^j \partial\theta^k} D_{\mathrm{KL}}(p_{\theta'} \| p_\theta)\Big|_{\theta'=\theta}$$

$$\Gamma^{(m)k}_{ij} = g^{kl}\Gamma^{(m)}_{ijl}$$

Dual Affine Connections and α-Connections

Define a one-parameter family of connections:

Amari's α-Connection

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(e)} + \frac{1-\alpha}{2}\nabla^{(m)}$$

- $\nabla^{(e)} = \nabla^{(1)}$ (exponential)
- $\nabla^{(m)} = \nabla^{(-1)}$ (mixture)
- $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metric

Duality

Connections $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual w.r.t. g:

$$X \cdot g(Y, Z) = g(\nabla_X^{(\alpha)} Y, Z) + g(Y, \nabla_X^{(-\alpha)} Z)$$

Interpretation in Exponential and Mixture Coordinates

Let $p(x;\theta)$ be an exponential family:

$$p(x;\theta) = \exp\left(\theta^i F_i(x) - \psi(\theta)\right)$$

- Exponential coordinates θ: $\nabla^{(e)}$-flat
- Expectation parameters $\eta_i = \mathbb{E}[F_i]$: $\nabla^{(m)}$-flat

Flatness: A coordinate system is flat w.r.t. a connection ∇ if all Christoffel symbols vanish in those coordinates.

This dual flatness underlies dually flat geometry and the existence of canonical divergences (e.g., KL).

Mirror Descent Dynamics in Geometry (I/II)

Figure: Optimization of $f(\theta) = (\theta - 2)^2$ using mirror descent

- Blue curve: target function $f(\theta)$
- Dashed gray: convex potential $\psi(\theta) = \frac{1}{2}\theta^2$
- Black path: updates via mirror descent using dual geometry

Update steps:

$$\begin{aligned} \eta_t &= \nabla\psi(\theta_t) \\ \eta_{t+1} &= \eta_t - \eta\nabla f(\theta_t) \\ \theta_{t+1} &= \nabla\psi^*(\eta_{t+1}) \end{aligned}$$

[Insert figure: Mirror descent update on $f(\theta)$]

Mirror Descent as Optimization in Dually Flat Geometry (II/II)

Mirror Descent = Natural Gradient Descent in Dual Space

Why it works:

- Uses structure from convex potential ψ
- Takes steps in dual space (expectation parameters)
- Returns to primal space (natural parameters) using Legendre dual

Dually flat manifold:

- θ-space: natural (e-flat), $\nabla^{(1)}$
- η-space: expectation (m-flat), $\nabla^{(-1)}$

Mirror descent traverses these dual geodesics.

[Re-show figure or zoom into geometric step]

Mirror Descent on Dually Flat Manifolds (I/II)

Optimization in flat space:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

uses Euclidean geometry.

Mirror Descent: Uses a Bregman divergence D_ψ from a convex potential ψ:

$$\begin{aligned} \eta_t &= \nabla \psi(x_t) \quad \text{(map to dual space)} \\ \eta_{t+1} &= \eta_t - \eta \nabla f(x_t) \quad \text{(gradient step)} \\ x_{t+1} &= \nabla \psi^*(\eta_{t+1}) \quad \text{(map back via Legendre dual)} \end{aligned}$$

Interpretation: Gradient steps happen in the **dual space** (m-coordinates), then map back to the primal (e-coordinates).

Geometry of Mirror Descent on Dually Flat Manifolds (II/II)

Dually flat structure:

- Flatness in θ (natural) space $\to \nabla^{(1)}$
- Flatness in η (expectation) space $\to \nabla^{(-1)}$
- Linked by $\eta = \nabla\psi(\theta)$ and $\theta = \nabla\psi^*(\eta)$

Mirror descent = natural gradient descent in dual geometry

Applications:

- Online learning (e.g., AdaGrad, exponentiated gradient)
- Variational inference updates
- Reinforcement learning (policy gradient with dual geometry)

Policy Gradient: Flat vs. Geometric Updates (I/III)

Goal: Optimize expected return by adjusting policy parameters

$$J(\theta) = \mathbb{E}_{\pi_\theta}[R]$$

where $\pi_\theta(a \mid s)$ is a parameterized stochastic policy.

Vanilla gradient ascent:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\theta_t)$$

Problem: This ignores the geometry of the policy space — can be unstable or slow.

Solution: Use **natural gradient** (mirror descent) — respects underlying geometry.

Natural Gradient as Mirror Descent (II/III)

Fisher Information Metric: Defines local geometry of policy space:

$$g_{ij}(\theta) = \mathbb{E}_{\pi_\theta}\left[\partial_i \log \pi_\theta(a \mid s)\, \partial_j \log \pi_\theta(a \mid s)\right]$$

Natural Gradient:

$$\tilde{\nabla}_\theta J = g^{-1}(\theta) \nabla_\theta J$$

This is equivalent to a mirror descent step on a dually flat manifold!

Primal = natural parameters (θ), Dual = expected features (η)

Update path follows m-geodesics in dual space, then maps back via ψ^*

Information-Geometric Policy Gradient (III/III)

Mirror descent view of policy updates:

- Choose convex potential $\psi(\theta)$ (e.g., KL divergence or entropy)
- Update in dual space: $\eta_{t+1} = \eta_t + \eta \cdot \text{advantage estimate}$
- Map back: $\theta_{t+1} = \nabla \psi^*(\eta_{t+1})$

Benefits:

- Adaptive learning via geometry
- Robustness in high-variance environments
- Theoretical guarantees from convex optimization

Used in: Trust Region Policy Optimization (TRPO), Natural Policy Gradient (NPG), and many modern RL algorithms.

Question?