Mini-Course on Information Geometry

Introduction

Herlock Rahimi
Department of Electrical and Computer Engineering
Yale University

June 5, 2025

Overview

Section 3.1: Duality of Connections

Motivation

- When studying the Fisher metric g and the α-connections $\nabla^{(\alpha)}$, we gain deeper insight by treating them as a triple:

$$(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$$

- These connections are not arbitrary but **dually coupled** through the metric g.
- Duality plays a central role in both theoretical and applied aspects of information geometry:
    - Symmetry in statistical models.
    - Unified geometric treatment.
    - Generalization of metric connections.

Definition of Dual Connections

Let S be a manifold with a Riemannian metric $g = \langle \cdot, \cdot \rangle$, and affine connections ∇ and ∇^*. They are **dual** with respect to g if for all vector fields $X, Y, Z \in \mathcal{T}(S)$,

$$Z\langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z^* Y \rangle$$

Notation: The triple (g, ∇, ∇^*) is called a *dualistic structure*.

Coordinate Expression of Duality

Let $\{\xi^i\}$ be local coordinates. Write the metric and connection coefficients as

$$g_{ij}, \quad \Gamma_{ij,k}, \quad \Gamma^*_{ij,k}$$

The duality condition becomes:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}$$

Implications:

- ∇^* is uniquely determined by g and ∇.
- $(\nabla^*)^* = \nabla$.
- The average $\frac{1}{2}(\nabla + \nabla^*)$ is a metric connection.

Theorem: Duality of α-connections

Theorem 3.1 (Amari):
For any statistical model, the α-connection and the $(-\alpha)$-connection are dual with respect to the Fisher metric.

$$\left(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)}\right) \text{ is a dualistic structure.}$$

Special Case:

- $\alpha = 1$: Exponential connection (**e-connection**).
- $\alpha = -1$: Mixture connection (**m-connection**).
- These are mutually dual.

Geometric Interpretation via Parallel Transport

Let $\gamma : [0,1] \to S$ be a smooth curve. Let $X(t)$, $Y(t)$ be vector fields along γ that are parallel:

$$\nabla_{\dot{\gamma}} X = 0, \quad \nabla^*_{\dot{\gamma}} Y = 0$$

Then the inner product is preserved:

$$\frac{d}{dt}\langle X(t), Y(t)\rangle = 0$$

Conclusion: Dual parallel transport preserves the inner product:

$$\langle \Pi_\gamma X, \Pi^*_\gamma Y\rangle = \langle X, Y\rangle$$

Curvature and Duality

Let R, R^* be the curvature tensors of ∇, ∇^*, respectively. Then:

$$\langle R(X,Y)Z, W\rangle = -\langle R^*(X,Y)W, Z\rangle$$

Hence,

$$R = 0 \quad \Leftrightarrow \quad R^* = 0$$

But: This symmetry does not extend to torsion tensors.

Section 3.2: Divergences and Contrast Functions

Why Divergences?

- In information geometry, **divergences** serve as generalized distance measures.
- Unlike metrics, they are typically asymmetric but still encode geometric information.
- They allow us to:
  - Induce a Riemannian metric (second-order info),
  - Define affine connections (third-order info),
  - Study statistical and inferential problems geometrically.
- Examples include the Kullback–Leibler divergence, Hellinger distance, and f-divergences.

Definition of a Divergence

Let S be a manifold. A function $D(p \parallel q) : S \times S \to \mathbb{R}$ is a **divergence** (or contrast function) if:

$$D(p \parallel q) \geq 0, \quad \text{and } D(p \parallel q) = 0 \iff p = q$$

Note: Unlike a metric, a divergence may not be symmetric and does not obey the triangle inequality. **Key idea:** Divergences provide a rich geometric structure through their Taylor expansions.

From Divergence to Metric

Let $D(p \parallel q)$ be smooth. Then we define the induced Riemannian metric $g^{(D)}$ by:

$$g_{ij}^{(D)}(p) := -\left.\frac{\partial^2 D(p \parallel q)}{\partial p^i \partial q^j}\right|_{q=p}$$

Equivalently, for vector fields $X, Y \in T_p(S)$:

$$\langle X, Y \rangle_D := -D[XY]$$

Interpretation: The divergence function determines how sharply D increases near p, i.e., the local curvature around the diagonal $p = q$.

From Divergence to Connection

Divergences also determine an affine connection $\nabla^{(D)}$ via:

$$\Gamma^{(D)}_{ij,k} := -D[\partial_i \partial_j \| \partial_k]$$

or equivalently:

$$\langle \nabla^{(D)}_X Y, Z \rangle_D = -D[XY \| Z]$$

Intuition: The divergence encodes not just distances but how these distances curve and change — captured by the third-order term.

Second and Third Order Expansion

Locally, we can expand the divergence as:

$$D(p \parallel q) = \frac{1}{2} g_{ij}^{(D)}(q)\Delta\xi^i\Delta\xi^j + \frac{1}{6} h_{ijk}^{(D)}(q)\Delta\xi^i\Delta\xi^j\Delta\xi^k + o(\|\Delta\xi\|^3)$$

where:

$$h_{ijk}^{(D)} = \partial_i g_{jk}^{(D)} + \Gamma_{jk,i}^{(D)}$$

Geometric Interpretation:

- 2nd-order term $\Rightarrow$ Riemannian structure (shape of “balls”).
- 3rd-order term $\Rightarrow$ affine structure (how “balls” twist).

Dual Connection from Dual Divergence

Define the dual divergence as:

$$D^*(p \parallel q) := D(q \parallel p)$$

Then the connection $\nabla^{(D^*)}$ is dual to $\nabla^{(D)}$ with respect to $g^{(D)}$. **Theorem 3.4:**

$$\left(g^{(D)}, \nabla^{(D)}, \nabla^{(D^*)}\right) \text{ is a dualistic structure}$$

Key Insight: Any smooth divergence gives rise to a dual pair of affine connections — this is the foundation of information geometry.

Example: f-Divergence (Csiszár)

Given a convex function $f : (0, \infty) \to \mathbb{R}$, define:

$$D_f(p \parallel q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx$$

Properties:

- General framework including Kullback–Leibler, Hellinger, and χ^2 divergences.
- Invariant under sufficient statistics.
- Duality: $D_f^*(p \parallel q) = D_{f^*}(p \parallel q)$, where $f^*(u) = uf(1/u)$.

Applications:

- Hypothesis testing
- Information bounds (e.g., Cramér–Rao)
- Robust estimation

The α-Divergence Family

The α-divergence is defined by:

$$D^{(\alpha)}(p \parallel q) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - \int p^{(1-\alpha)/2} q^{(1+\alpha)/2}\right) & \alpha \neq \pm 1 \\ \int p(x) \log \frac{p(x)}{q(x)} dx & \alpha = 1 \quad \text{(Kullback)} \\ \int q(x) \log \frac{q(x)}{p(x)} dx & \alpha = -1 \end{cases}$$

Key:

$$D^{(\alpha)}(p \parallel q) = D^{(-\alpha)}(q \parallel p)$$

This divergence induces:

$$(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$$

— the canonical dualistic structure of statistical models.

Why This Matters

- Divergences generalize distance in contexts where true distances (metrics) don't exist — e.g., probability distributions.
- They allow geometry to emerge from statistical notions:
  - Fisher metric from Kullback–Leibler.
  - e-/m-connections from forward/reverse KL.
- **Core insight:** Statistical inference, estimation, and hypothesis testing can be understood geometrically using divergences and their induced structures.

Section 3.3: Dually Flat Spaces

Why Dually Flat Spaces?

- When both connections ∇ and ∇^* are flat and dual w.r.t. a metric g, the manifold has rich geometric and analytical structure.
- These spaces allow:
  - Affine coordinate systems for both connections.
  - A natural potential function (convex), enabling Legendre duality.
  - Efficient geometric reasoning in information theory and statistics.
- Common in exponential families and mixture families.

Definition: Dually Flat Space

A manifold S with metric g and connections ∇, ∇^* is called a **dually flat space** if:

(S, g, ∇, ∇^*) is a dualistic structure and ∇, ∇^* are both flat.

Key consequence: There exist coordinate systems $[\theta^i]$ and $[\eta^j]$ such that:

$$\langle \partial_{\theta^i}, \partial_{\eta^j} \rangle = \delta_i^j$$

We say these coordinate systems are **mutually dual**.

Geometry of Flatness

In a dually flat space:

- ∇-geodesics are straight lines in θ-coordinates.
- ∇^*-geodesics are straight in η-coordinates.
- The metric g connects the two via:

$$\frac{\partial \eta_j}{\partial \theta^i} = g_{ij}, \quad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij}$$

- Dual flatness $\Rightarrow$ integrability of coordinate duality.

Result: Dual coordinate systems give complete control over geometry via linear algebra and convex analysis.

Potentials and Legendre Duality

There exists a strictly convex function $\psi(\theta)$ such that:

$$\eta_i = \frac{\partial \psi}{\partial \theta^i}, \quad g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$$

Define the Legendre dual:

$$\varphi(\eta) := \sup_{\theta} \left(\theta^i \eta_i - \psi(\theta)\right)$$

Then:

$$\theta^i = \frac{\partial \varphi}{\partial \eta_i}, \quad g^{ij} = \frac{\partial^2 \varphi}{\partial \eta_i \partial \eta_j}$$

Interpretation: Convex duality between parameters and expectations.

Mutually Dual Coordinates

Theorem (3.6): Let $[\theta^i]$ be ∇-affine coordinates. Then:

- There exists a unique ∇^*-affine dual system $[\eta_i]$.
- They satisfy:

$$\langle \partial_{\theta^i}, \partial_{\eta^j} \rangle = \delta_i^j$$

- They are linked via Legendre transformation between ψ and φ.

Example: In exponential families,

$$\theta_i = \text{natural parameters}, \quad \eta_i = \mathbb{E}_\theta[F_i]$$

Canonical Divergence on Dually Flat Space

Given potentials $\psi(\theta)$ and $\varphi(\eta)$, define:

$$D(p \parallel q) := \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q)$$

Then:

- $D(p \parallel q) \geq 0$ with equality iff $p = q$
- Induces g, ∇, ∇^*
- Is asymmetric: $D(q \parallel p) = D^*(p \parallel q)$

This is the canonical divergence.

The Generalized Pythagorean Theorem

Let γ_1 be a ∇-geodesic from $p \to q$, and γ_2 a ∇^*-geodesic from $q \to r$. If $\gamma_1 \perp \gamma_2$ at q, then:

$$D(p \parallel r) = D(p \parallel q) + D(q \parallel r)$$

Interpretation:

- The divergence plays the role of squared distance.
- Geodesic orthogonality additive divergence.
- Extends projection and approximation to the information geometric setting.

Let $M \subset S$ be a ∇^*-flat submanifold. Then:

$$\arg\min_{q \in M} D(p \parallel q)$$

is characterized by:

- The minimizing $q \in M$ satisfies that the ∇-geodesic from p to q is orthogonal to M.
- This is called the ∇-**projection** of p onto M.

Important in:

- Exponential family model fitting
- Variational inference
- Bregman projections in machine learning

Comparison to Classical Riemannian Geometry

- In Riemannian geometry:
$$D(p \parallel q) = \frac{1}{2}\|p - q\|^2$$
and all geodesics coincide.
- In information geometry:
$$D(p \parallel q) \neq D(q \parallel p)$$
and two different geodesics connect p and q: one for ∇, one for ∇^*.
- The divergence function replaces the role of squared distance.

Conclusion: Information geometry generalizes Riemannian geometry by accommodating statistical asymmetry.

Exponential Family Form of Normal Distribution

Consider the family of univariate normal distributions with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$:

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

This can be written in exponential family form:

$$p(x;\theta) = \exp\left(\theta^1 x + \theta^2 x^2 - \psi(\theta) + C(x)\right)$$

where:

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

Computing $\psi(\theta)$: Step-by-Step

- We ensure normalization:

$$\int \exp\left(\theta^1 x + \theta^2 x^2\right) dx = e^{\psi(\theta)}$$

- Complete the square:

$$\theta^1 x + \theta^2 x^2 = \theta^2 \left(x^2 + \frac{\theta^1}{\theta^2} x \right)$$

$$= \theta^2 \left[\left(x + \frac{\theta^1}{2\theta^2} \right)^2 - \left(\frac{\theta^1}{2\theta^2} \right)^2 \right]$$

- Therefore:

$$\int \exp(\theta^1 x + \theta^2 x^2) dx = \exp\left(-\frac{(\theta^1)^2}{4\theta^2} \right) \int \exp\left(\theta^2 \left(x + \frac{\theta^1}{2\theta^2} \right)^2 \right) dx$$

- Let $y = x + \frac{\theta^1}{2\theta^2}$. Then:

$$\int e^{\theta^2 y^2} dy = \sqrt{\frac{\pi}{-\theta^2}} \quad (\text{if } \theta^2 < 0)$$

Expectation Coordinates η

General case: Expectation parameters are:

$$\eta_i = \mathbb{E}_\theta[F_i(x)] = \frac{\partial\psi}{\partial\theta^i}$$

In our case:

$$F_1(x) = x, \quad F_2(x) = x^2$$

So:

$$\eta_1 = \mu, \quad \eta_2 = \mu^2 + \sigma^2 = \mathbb{E}[X^2]$$

Derivatives of $\psi(\theta)$

From:

$$\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta^2}\right)$$

We compute:

$$\eta_1 = \frac{\partial \psi}{\partial \theta^1} = -\frac{\theta^1}{2\theta^2} = \mu$$

$$\eta_2 = \frac{\partial \psi}{\partial \theta^2} = \frac{(\theta^1)^2}{4(\theta^2)^2} - \frac{1}{2\theta^2} = \mu^2 + \sigma^2$$

Thus, we recover the expectation coordinates!

Fisher Metric from ψ

Compute the second derivatives:

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$$

which gives the Fisher information matrix. In this normal family:

- It measures sensitivity of log-likelihood.
- Provides the Riemannian metric for the dually flat structure.

General Formulation:

- *e*-geodesic: linear interpolation in θ

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

- *m*-geodesic: linear interpolation in η

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

In our case: These represent:

- Natural parameter space geodesics $\Rightarrow$ exponential updates
- Expectation space geodesics $\Rightarrow$ moment-based interpolation

Canonical Divergence: General Form

For any dually flat space:

$$D(p \parallel q) = \psi(\theta_q) + \varphi(\eta_p) - \theta_q^i \eta_{p,i}$$

This is a Bregman divergence induced by ψ.

KL Divergence for Normals: Step-by-Step

Let:

$$p = \mathcal{N}(\mu_1, \sigma_1^2), \quad q = \mathcal{N}(\mu_2, \sigma_2^2)$$

Then the KL divergence is:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Compute:

$$\log \frac{p(x)}{q(x)} = \log \left(\frac{\sigma_2}{\sigma_1} \right) + \frac{(x - \mu_2)^2}{2\sigma_2^2} - \frac{(x - \mu_1)^2}{2\sigma_1^2}$$

Take expectation over p:

$$\mathbb{E}_p[x] = \mu_1, \quad \mathbb{E}_p[x^2] = \mu_1^2 + \sigma_1^2$$

Final result:

$$D_{\mathrm{KL}}(p \parallel q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

KL as Canonical Divergence in θ and η

The expression:

$$D(p \parallel q) = \psi(\theta_q) + \varphi(\eta_p) - \theta_q^i \eta_{p,i}$$

when computed using the above ψ, θ, and η, gives the exact KL divergence:

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \sigma_1^2) \parallel \mathcal{N}(\mu_2, \sigma_2^2))$$

Thus, KL is the canonical divergence for this exponential family.

Section 3.4: Canonical Divergence

Why Canonical Divergence?

- Many divergences can induce the same dualistic structure (g, ∇, ∇^*).
- Question: Is there a unique divergence naturally tied to a dually flat space?
- Answer: **Yes — the canonical divergence**, constructed directly from:
  - Dual coordinate systems θ, η,
  - Convex potentials ψ, φ,
  - Their Legendre relation.
- This divergence underlies much of the theory and applications in information geometry.

Definition: Canonical Divergence

Let (θ^i) be ∇-affine coordinates, (η_i) the ∇^*-affine dual coordinates. Define the potentials:

$$\psi(\theta) := \int \eta_i d\theta^i, \quad \varphi(\eta) := \sup_\theta \left(\theta^i \eta_i - \psi(\theta)\right)$$

The **canonical divergence** is:

$$D(p \parallel q) := \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q)$$

Key: It depends only on the dually flat structure, not on extraneous data.

Properties of Canonical Divergence

- $D(p \parallel q) \geq 0$, with equality iff $p = q$
- Asymmetric: $D(p \parallel q) \neq D(q \parallel p)$
- Induces:

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} = \frac{\partial^2 \varphi}{\partial \eta_i \partial \eta_j}$$

- Connections:

$$\nabla := \nabla^{(D)}, \quad \nabla^* := \nabla^{(D^*)}$$

- Coordinate independence: any change preserving dual flatness keeps D invariant.

Geometric Insight

- $D(p \parallel q)$ measures discrepancy between two distributions in terms of potentials:

$$\text{Divergence} = \text{energy at } p + \text{energy at } q - \text{coupling term}$$

- When $\psi = \frac{1}{2}\|\theta\|^2$, this reduces to:

$$D(p \parallel q) = \frac{1}{2}\|\theta(p) - \theta(q)\|^2$$

- So D generalizes squared distance in Euclidean space.

Key: It's intrinsic to the manifold structure — not arbitrarily chosen.

Theorem 3.7: Characterization of Canonical Divergence

A divergence D is canonical if and only if for all p, q, r:

$$D(p \parallel q) + D(q \parallel r) - D(p \parallel r) = \left(\theta^i(p) - \theta^i(q)\right)\left(\eta_i(r) - \eta_i(q)\right)$$

Interpretation:

- The divergence satisfies a generalized triangle equality.
- Expresses how divergence decomposes along dual geodesics.

Theorem 3.8: Pythagorean Relation

Let:

- γ_1: ∇-geodesic from $p \to q$,
- γ_2: ∇^*-geodesic from $q \to r$,
- and $\gamma_1 \perp \gamma_2$ at q,

Then:

$$D(p \parallel r) = D(p \parallel q) + D(q \parallel r)$$

Consequences:

- Enables orthogonal decomposition of divergence.
- Used in statistical projection and optimization.

Projection Theorem and Optimization

Let $M \subset S$ be ∇^*-flat. For a given $p \in S$, the point

$$q^* = \arg\min_{q \in M} D(p \parallel q)$$

is characterized by:

- The ∇-geodesic $p \to q^*$ is orthogonal to M.
- The divergence is minimized in the statistical sense.

Applications:

- Maximum likelihood estimation.
- Information projection.
- Variational approximation.

Curve Divergence: A Local Perspective

Let $\gamma : [a, b] \to S$ be a smooth curve. Define the divergence along γ as:

$$D(\gamma) := D_\gamma(\gamma(b) \parallel \gamma(a))$$

When γ is 1D, it is always dually flat, so canonical divergence exists along any curve. If γ is a ∇-geodesic:

$$D(\gamma) = \int_a^b (b - s) \cdot g_\gamma(s) ds$$

Generalizes: Curve length $\Rightarrow$ curve divergence.

Section 3.5: Dualistic Structure of Exponential Families

Why Focus on Exponential Families?

- They are central to statistics: maximum entropy, sufficient statistics, MLEs, conjugacy.
- In information geometry, they form canonical examples of **dually flat spaces**.
- Both natural θ and expectation η parameters form affine coordinate systems:

$$\theta \rightsquigarrow \nabla^{(1)}\text{-affine}, \quad \eta \rightsquigarrow \nabla^{(-1)}\text{-affine}$$

- The Fisher metric and KL divergence naturally arise from their structure.

Exponential Family Form

An exponential family over a space $\mathcal{X}$ is given by:

$$p(x;\theta) = \exp\left[\sum_i \theta^i F_i(x) - \psi(\theta) + C(x)\right]$$

Where:

- θ^i — natural parameters
- $F_i(x)$ — sufficient statistics
- $\psi(\theta)$ — log-partition function (convex)
- $C(x)$ — base measure adjustment

Key: The geometry of $\theta \mapsto p(x;\theta)$ is governed by ψ.

Dual Coordinate Systems

Define:

$$\eta_i = \mathbb{E}_\theta[F_i(x)] = \frac{\partial \psi}{\partial \theta^i}$$

Then:

$$\theta \text{ is } \nabla^{(1)}\text{-affine}, \quad \eta \text{ is } \nabla^{(-1)}\text{-affine}$$

Why dual?

$$\frac{\partial \eta_j}{\partial \theta^i} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} = g_{ij}$$

So θ, η are related by Legendre transformation:

$$\varphi(\eta) := \theta^i \eta_i - \psi(\theta)$$

Example: Normal Distribution (Univariate)

Canonical form:

$$p(x; \theta) = \exp\left[\theta_1 x + \theta_2 x^2 - \psi(\theta)\right]$$

where:

$$\theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = -\frac{1}{2\sigma^2}$$

Dual coordinates:

$$\eta_1 = \mathbb{E}[X] = \mu, \quad \eta_2 = \mathbb{E}[X^2] = \mu^2 + \sigma^2$$

Summary: θ and η parameterize the same family with dual geometric structure.

Fisher Metric in Exponential Families

The Fisher information is:

$$g_{ij}(\theta) = \mathbb{E}_\theta[(F_i - \eta_i)(F_j - \eta_j)]$$

But also:

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$$

and

$$g^{ij} = \frac{\partial^2 \varphi}{\partial \eta_i \partial \eta_j}$$

Result: (θ, η) coordinate systems induce dual representations of the Fisher geometry.

KL Divergence as Canonical Divergence

For exponential families:

$$D(p_\theta \parallel p_{\theta'}) = \psi(\theta') - \psi(\theta) - (\theta' - \theta)^i \eta_i$$

This is the **Kullback–Leibler divergence**. **Therefore:**

$$D^{(1)}(p_\theta \parallel p_{\theta'}) = D_{\mathrm{KL}}(p_{\theta'} \parallel p_\theta)$$

$$D^{(-1)}(p_\theta \parallel p_{\theta'}) = D_{\mathrm{KL}}(p_\theta \parallel p_{\theta'})$$

Canonical divergence coincides with KL divergence — a deep unification!

Efficient Estimators and Duality

Define $\hat{\eta}_i(x) := F_i(x)$. Then:

- $\hat{\eta}$ is an unbiased estimator of η.
- Covariance matrix is:

$$\mathrm{Cov}_\theta[\hat{\eta}_i, \hat{\eta}_j] = g_{ij}$$

- Hence, $\hat{\eta}$ achieves the Cramér–Rao bound:

$$\mathrm{Var}(\hat{\eta}) \geq g^{-1}$$

Theorem (3.12): A coordinate system admits an efficient estimator iff it is $\nabla^{(-1)}$-affine (i.e., expectation parameters).

Section 3.8: Statistical Estimation and Dual Flatness

Why Geometry for Estimation?

- Estimators are functions from data to parameters: $\hat{\theta}(x)$
- We want estimators to be:
  - Unbiased,
  - Efficient (minimum variance),
  - Invariant under transformations.
- Information geometry gives a **geometric criterion** for optimality.
- It characterizes efficient estimators via flatness and orthogonality in the statistical manifold.

Definition: Unbiased Estimator

Let ξ^i be local coordinates on a manifold S. An estimator $\hat{\xi}^i(x)$ is **unbiased** if:

$$\mathbb{E}_\xi[\hat{\xi}^i(x)] = \xi^i$$

Example: In exponential families, $\hat{\eta}_i(x) := F_i(x)$ is an unbiased estimator of η_i.

Score Function and Fisher Information

Define:

$$l_i(x;\xi) := \frac{\partial}{\partial \xi^i} \log p(x;\xi)$$

Then the **Fisher information matrix** is:

$$g_{ij} := \mathbb{E}_{\xi}[l_i l_j]$$

This acts as the metric in the statistical manifold — it determines the curvature of likelihood.

Covariance Matrix of Estimator

Define the covariance matrix of an estimator $\hat{\xi}$:

$$V^{ij} := \mathbb{E}_{\xi}\left[(\hat{\xi}^i - \xi^i)(\hat{\xi}^j - \xi^j)\right]$$

Then the **Cramér–Rao inequality** holds:

$$V \geq g^{-1}$$

Equality estimator achieves minimum variance — this is our geometric goal.

Theorem 3.14: Dual Flatness and Efficient Estimation

Statement:

- A coordinate system $[\xi^i]$ admits an efficient estimator iff it is $\nabla^{(-1)}$-affine.
- Equivalently: the coordinate curves are $\nabla^{(-1)}$-geodesics.

Implication: Geometry of the manifold constrains what can be efficiently estimated.

Example: Exponential Family Revisited

- θ^i — natural parameters (not efficiently estimable).
- $\eta_i = \mathbb{E}_\theta[F_i]$ — expectation parameters.
- Estimator: $\hat{\eta}_i(x) = F_i(x)$.
- Covariance:

$$\mathrm{Cov}[\hat{\eta}_i, \hat{\eta}_j] = g_{ij}$$

- Hence: $\hat{\eta}$ saturates Cramér–Rao bound efficient.

Geometric Interpretation

- Estimation corresponds to projection of the empirical distribution.
- Efficient estimation occurs when projection is orthogonal in dual geometry.
- Dual flatness affine coordinate systems linear unbiased estimators.
- The curvature of the manifold limits achievable variance.

Summary: Estimation and Geometry

- Information geometry provides:
  - A criterion for efficiency: dual flatness.
  - A metric to measure variance: Fisher information.
  - A way to construct efficient estimators: dual coordinates.
- Estimation becomes not just a numerical problem — but a geometric one.

Insight: Geometry governs optimality.

Question?