Information Geometry & Neuroscience

A reading course by Aniket Deshpande
with Matthew Singh

Contents

Preface

These notes were compiled during a reading course I completed in the fall of 2025 on *information geometry*, advised by Dr. Matthew Singh.

The field of information geometry traces back to the work of C.R. Rao (see [Rao92]), where the Fisher matrix was first used as a Riemannian metric. The theories were modernized by Shun'ichi Amari [Ama83], and an abundance of work has been done to find geometric structure in information theory and machine learning.

Information geometry is a study of statistical manifolds. Today, the field has emerged as highly interdisciplinary with applications in signal processing, deep learning, optimal transport, neuroscience, and quantum information. This reading course (and these notes) will focus on geometric aspects of information, control, and coding theory with an emphasis on applications to neuroscience.

These notes will follow a multitude of papers and foundational references.

these notes are a work in progress. they will be updated throughout the semester.

1 Foundations of Information Geometry

We begin with a survey of the mathematical foundations of information geometry. A finer definition of the field is the study of the *geometry of decision making* [Nie20].

1.1 Some Geometry

Informally, a *smooth* D-dimensional manifold $\mathcal{M}$ is a topological space that locally behaves like D-dimensional Euclidean space $\mathbb{R}^D$.

Definition 1 (Manifolds). A topological space $\mathcal{M}$ is a D-dimensional *manifold* if for every point $p \in \mathcal{M}$, there exists an open neighborhood $U \subseteq \mathcal{M}$ containing p and a homeomorphism $\phi : U \to V$ where V is an open subset of $\mathbb{R}^D$. The pair (U, ϕ) is called a *chart*, and the collection of charts that cover $\mathcal{M}$ is called an *atlas*. If the transition maps between overlapping charts are smooth (infinitely differentiable), then $\mathcal{M}$ is called a *smooth manifold*.

Geometric objects, i.e. points, balls, and vector fields, and entities, i.e. functions and operators, live on a manifold $\mathcal{M}$. These objects are *coordinate-free*, meaning that they do not depend on the choice of chart. However, we can express these objects in *any* local coordinate system of an atlas $\mathcal{A} = \{(U_i, x_i)\}_i$ of charts (U_i, x_i) where $x_i : U_i \to V_i \subseteq \mathbb{R}^D$ is a homeomorphism. In information geometry, we generally handle a single chart fully covering the manifold $\mathcal{M}$, and so we will often drop the subscript i and simply write (U, x).

A C^k manifold is obtained when the change of chart transformations are also C^k. These manifolds are *smooth* when $k = \infty$. At each point $p \in \mathcal{M}$, we can define a *tangent space*.

Definition 2 (Tangent Spaces). At each point $p \in \mathcal{M}$, where $\mathcal{M}$ is a D-dimensional smooth manifold, the *tangent space* $T_p\mathcal{M}$ is a D-dimensional vector space that consists of the tangent vectors to all possible curves on $\mathcal{M}$ passing through p. Formally, if $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ is a smooth curve with $\gamma(0) = p$, then the tangent vector to γ at p is defined as the equivalence class of curves that have the same derivative at p. The collection of all such tangent vectors forms the tangent space $T_p\mathcal{M}$.

Example 1. Consider the 2-dimensional sphere $\mathbb{S}^2$ embedded in $\mathbb{R}^3$. At any point p on the sphere, the tangent space $T_p\mathbb{S}^2$ is a 2-dimensional plane that touches the sphere at p and is perpendicular to the radius vector from the origin to p. This tangent space contains all possible directions in which one can move tangentially from the point p on the surface of the sphere.

On any smooth manifold, we can define two *independent* structures, the *metric tensor* g and an *affine connection* ∇.

Definition 3 (Metric Tensors). A *metric tensor* g on a smooth manifold $\mathcal{M}$ is a smooth, symmetric, positive-definite bilinear form that assigns to each point $p \in \mathcal{M}$ an inner product $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ on the tangent space $T_p\mathcal{M}$.

This inner product allows us to measure lengths of tangent vectors and angles between them, thereby defining a notion of distance and geometry on the manifold, vital for measuring statistical distances.

Definition 4 (Affine Connections). An *affine connection* ∇ is a differential operator allows one to define the following.

1. A *covariant derivative operator*: For any two smooth vector fields X and Y on $\mathcal{M}$, the covariant derivative $\nabla_X Y$ is a new vector field that represents the rate of change of Y in the direction of X.
2. *Parallel transport*: Π_c^∇ defines a way to transport vectors between tangent planes along any smooth curve $c : [0,1] \to \mathcal{M}$ while preserving the notion of "parallelism" defined by the connection ∇.
3. The notion of ∇-*geodesics* γ_∇ define autoparallel curves, extending the idea of "straight lines" to curved spaces. A curve $\gamma : [0,1] \to \mathcal{M}$ is a geodesic if its tangent vector field $\dot{\gamma}(t)$ is parallel transported along itself, i.e., $\nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) = 0$ for all $t \in [0,1]$.

The *tangent bundle* of $\mathcal{M}$ can be defined as a sort of "union" of all tangent spaces.

$$T\mathcal{M} := \bigcup_{p \in \mathcal{M}} T_p\mathcal{M} = \{(p,v) : p \in \mathcal{M}, v \in T_p\mathcal{M}\}. \tag{1}$$

This means that $T\mathcal{M}$ itself is a $2D$-dimensional manifold. [1] Tangent vectors v play the role of directional derivatives, with vf informally meaning the derivative of a smooth function $f \in \mathfrak{F}(\mathcal{M})$ along the direction v. Since the manifolds are abstracted and not embedded in some Euclidean space, we cannot view the vector as an "arrow" anchored on the manifold. Instead, vectors can be understood as directional derivatives or equivalence classes of smooth curves at a point.

Definition 5 (*Smooth* Vector Fields). A smooth vector field X is defined as a "cross-section" of the tangent bundle. $X \in \mathfrak{X}(\mathcal{M}) = \Gamma(T\mathcal{M})$ is a smooth vector field if $X : \mathcal{M} \to T\mathcal{M}$ such that $\pi \circ X = \mathrm{id}_\mathcal{M}$ where $\pi : T\mathcal{M} \to \mathcal{M}$ is the natural projection map $\pi(p,v) = p$. In other words, X assigns to each point $p \in \mathcal{M}$ a tangent vector $X_p \in T_p\mathcal{M}$ smoothly.

In local coordinates $(x^1, \dots, x^D)$ on a chart (U,x), a vector field X can be expressed as $X = X^i \frac{\partial}{\partial x^i}$, where we use the Einstein summation convention (summing over repeated indices). The coordinate basis vectors $\left\{\frac{\partial}{\partial x^i}\right\}_{i=1}^D$ form a local basis for the tangent space at each point.

Definition 6 (Cotangent Spaces and Differential Forms). At each point $p \in \mathcal{M}$, the *cotangent space* $T_p^*\mathcal{M}$ is the dual vector space to the tangent space $T_p\mathcal{M}$. Elements of $T_p^*\mathcal{M}$, called *covectors* or *one-forms*, are linear functionals $\omega : T_p\mathcal{M} \to \mathbb{R}$. The *cotangent bundle* is defined as

$$T^*\mathcal{M} := \bigcup_{p \in \mathcal{M}} T_p^*\mathcal{M} = \left\{(p,\omega) : p \in \mathcal{M}, \omega \in T_p^*\mathcal{M}\right\}. \tag{2}$$

A smooth *differential 1-form* is a smooth section of the cotangent bundle, i.e., a smooth map $\omega : \mathcal{M} \to T^*\mathcal{M}$ that assigns to each point p a covector $\omega_p \in T_p^*\mathcal{M}$.

In local coordinates, a differential 1-form can be expressed as $\omega = \omega_i dx^i$, where $\{dx^i\}_{i=1}^D$ forms the dual basis to $\left\{\frac{\partial}{\partial x^i}\right\}_{i=1}^D$, satisfying $dx^i\left(\frac{\partial}{\partial x^j}\right) = \delta^i_j$ (the Kronecker delta).

[1] The tangent bundle is a specific example of a *fiber bundle* with base manifold $\mathcal{M}$.

Definition 7 (Riemannian Manifolds). A *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a metric tensor g. The metric tensor in local coordinates can be written as $g = g_{ij} dx^i \otimes dx^j$, where $g_{ij}(p) = g_p\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$ are the components of the metric tensor at point p.

The metric tensor enables us to compute the length of a curve $\gamma : [a,b] \to \mathcal{M}$ as

$$L(\gamma) = \int_a^b \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\, dt, \tag{3}$$

and to define the *Riemannian distance* between two points $p, q \in \mathcal{M}$ as the infimum of lengths over all curves connecting p and q.

Definition 8 (Christoffel Symbols). Given a metric tensor g on a manifold $\mathcal{M}$, the *Christoffel symbols* Γ^k_{ij} of the *Levi-Civita connection* $\nabla^{(g)}$ are defined in local coordinates by

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\frac{\partial g_{il}}{\partial x^j} + \frac{\partial g_{jl}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l} \right), \tag{4}$$

where g^{kl} are the components of the inverse metric tensor satisfying $g^{kl} g_{lm} = \delta^k_m$. The covariant derivative of a vector field $Y = Y^j \frac{\partial}{\partial x^j}$ in the direction of a vector field $X = X^i \frac{\partial}{\partial x^i}$ is given by

$$\nabla^{(g)}_X Y = X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma^k_{ij} Y^j \right) \frac{\partial}{\partial x^k}. \tag{5}$$

The Levi-Civita connection is the unique torsion-free connection compatible with the metric, meaning that it preserves the metric under parallel transport: $\nabla^{(g)}_X g = 0$ for all vector fields X.

Definition 9 (Curvature Tensor). The *Riemann curvature tensor* R measures the non-commutativity of covariant differentiation and the failure of parallel transport to be path-independent. For vector fields $X, Y, Z \in \mathfrak{X}(\mathcal{M})$, the curvature tensor is defined as

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \tag{6}$$

where $[X,Y] = XY - YX$ is the Lie bracket of vector fields. In local coordinates, the components of the Riemann curvature tensor are

$$R^l_{ijk} = \frac{\partial \Gamma^l_{ik}}{\partial x^j} - \frac{\partial \Gamma^l_{ij}}{\partial x^k} + \Gamma^m_{ij} \Gamma^l_{mk} - \Gamma^m_{ik} \Gamma^l_{mj}. \tag{7}$$

A manifold is said to be *flat* if the Riemann curvature tensor vanishes everywhere. Flat manifolds locally resemble Euclidean space, and parallel transport is path-independent. The curvature tensor encodes fundamental information about the intrinsic geometry of the manifold, distinguishing curved spaces from flat ones.

Definition 10 (Ricci Curvature and Scalar Curvature). By contracting the Riemann curvature tensor, we obtain two important curvature measures:

1. The *Ricci curvature tensor* is obtained by contracting the first and third indices: $\mathrm{Ric}_{ij} = R^k_{ikj} = g^{kl}R_{kilj}$.
2. The *scalar curvature* is the trace of the Ricci tensor: $S = g^{ij}\mathrm{Ric}_{ij}$.

These quantities measure how the volume of geodesic balls differs from Euclidean balls, providing global information about the manifold's geometry.

By definition, an affine connection ∇ is said to be *metric compatible* with g when it satisfies for any triple (X,Y,Z) of vector fields the condition

$$X\langle Y,Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \tag{8}$$

which can be written as

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z). \tag{9}$$

Theorem 1 (Levi-Civita Connection). There exists a unique torsion-free affine connection compatible with them metric called the *Levi-Civita connection* $\nabla^{(g)}$.

The Christoffel symbols of the Levi-Civita connection can be expressed from the metric tensor g as follows

$$\Gamma^k_{ij} = \frac{1}{2}g^{kl}\left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\right), \tag{10}$$

where g^{kl} denote the matrix elements of the inverse of the metric tensor g_{ij}.

1.2 Introducing Information Manifolds

In information geometry, we consider two conjugate affine connections ∇ and ∇^* that are coupled to the metric g. The structure is written as $(\mathcal{M}, g, \nabla, \nabla^*)$. The important property is that these conjugate connections are metric compatible, therefore the induced dual parallel transport preserves the metric:

$$\langle u, v\rangle_{c(0)} = \left\langle \Pi^{\nabla}_{c(0)\to c(t)} u, \Pi^{\nabla^*}_{c(0)\to c(t)} v \right\rangle_{c(t)}. \tag{11}$$

We begin with a definiton.

Definition 11 (Conjugate Connections). A connection ∇^* is said to be *conjugate* (or dual) to a connection ∇ with respect to the metric tensor g if and only if, for any triple (X,Y,Z) of smooth vector fields, the following identity is satisfied:

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z). \tag{12}$$

We can check that the right-hand side is a scalar and that the left-hand side is a directional derivative of a real-valued function, another scalar. [2]

Definition 12 (Conjugate Connection Manifolds). The structure of a conjugate connection manifold (CCM) is denoted by $(\mathcal{M}, g, \nabla, \nabla^*)$, where (∇, ∇^*) is a pair of conjugate connections with respect to the metric tensor g on the smooth manifold $\mathcal{M}$.

[2] Conjugation is an involution: $(\nabla^*)^* = \nabla$.

A very important property of such structure is that the dual parallel transports of vectors preserves the metric. That is, for any smooth curve $c : [0,1] \to \mathcal{M}$, the inner product of two vectors $u, v \in T_{c(0)}\mathcal{M}$ is preserved under the primal transport on u, $\Pi_c^{\nabla} u$, and the dual transport on v, $\Pi_c^{\nabla^*} v$.

$$\langle u, v\rangle_{c(0)} = \left\langle \Pi^{\nabla}_{c(0)\to c(t)} u, \Pi^{\nabla^*}_{c(0)\to c(t)} v \right\rangle_{c(t)}. \tag{13}$$

In fact, the existence of one connection ∇ automatically induces the existence of a unique conjugate connection ∇^*.

Theorem 2. Given a connection ∇ on $(\mathcal{M}, g)$, there exists a unique conjugate connection ∇^* with respect to the metric tensor g.

Given a manifold equipped with a metric tensor g and a connection ∇, we can always build the conjugate connection ∇^* and define a *mean* connection

$$\bar{\nabla} = \frac{\nabla + \nabla^*}{2}.$$

with corresponding Christoffel symbols $\bar{\Gamma}$.

Theorem 3. The mean connection $\bar{\nabla}$ is self-conjugate, i.e. $\bar{\nabla} = \bar{\nabla}^*$, and coincides with the Levi-Civita connection.

$$\bar{\nabla} = \nabla^{\mathrm{LC}}. \tag{14}$$

The term "statistical manifold" was originally purely a geometric construction. It was later shown that we can always find a *statistical model* $\mathcal{P}$ corresponding to a given statistical manifold, and how we can derive an infinite family of CCMs from said manifold. Consider a *totally symmetric* [3] cubic $(0,3)$-tensor C on $(\mathcal{M}, g)$.

Definition 13 (Statistical Manifolds). A *statistical manifold* $(\mathcal{M}, g, C)$ is a Riemannian manifold $(\mathcal{M}, g)$ equipped with a totally symmetric cubic tensor C called the *Amari-Chentsov tensor* or *cubic form*. In local coordinates, the cubic tensor can be written as $C = C_{ijk} dx^i \otimes dx^j \otimes dx^k$, where C_{ijk} are symmetric in all three indices.

From a statistical manifold $(\mathcal{M}, g, C)$, we can construct a pair of conjugate connections (∇, ∇^*) such that

$$C_{ijk} = \Gamma_{ij,k} - \Gamma^*_{ij,k}, \tag{15}$$

where $\Gamma_{ij,k} = g_{kl}\Gamma^l_{ij}$ and $\Gamma^*_{ij,k} = g_{kl}(\Gamma^*)^l_{ij}$ are the Christoffel symbols of the first kind for ∇ and ∇^*, respectively. Conversely, given a CCM $(\mathcal{M}, g, \nabla, \nabla^*)$, we can define the cubic tensor C by the above relation.

1.3 The Fundamental Theorem of Information Geometry

The following theorem establishes the fundamental relationship between statistical manifolds and conjugate connection manifolds, and provides the foundation for constructing α-connections.

[3] That is, $C_{ijk} = C_{\sigma(i)\sigma(j)\sigma(k)}$ for any permutation σ.

Theorem 4 (Fundamental Theorem of Information Geometry). Given a statistical manifold $(\mathcal{M}, g, C)$, there exists a one-parameter family of conjugate connection manifolds $\{(\mathcal{M}, g, \nabla^{-\alpha}, \nabla^{\alpha})\}_{\alpha\in\mathbb{R}}$ such that:

1. For $\alpha = 0$, we have $\nabla^0 = \nabla^{\mathrm{LC}}$, the Levi-Civita connection.
2. The cubic tensor is related to the connections by

$$C_{ijk} = \Gamma^{(-\alpha)}_{ij,k} - \Gamma^{(\alpha)}_{ij,k}, \tag{16}$$

   where $\Gamma^{(\alpha)}_{ij,k}$ are the Christoffel symbols of the first kind for ∇^{α}.
3. The α-connections satisfy the relation

$$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{\alpha}{2} C_{ijk}, \tag{17}$$

   where $\Gamma^{(0)}_{ij,k}$ are the Christoffel symbols of the Levi-Civita connection.

This theorem shows that from any statistical manifold, we can construct an entire family of dualistic structures parameterized by $\alpha \in \mathbb{R}$. The parameter α interpolates between the two conjugate connections, with $\alpha = 0$ corresponding to the Levi-Civita connection.

1.4 Statistical Models and the Fisher Information Metric

We now connect the abstract geometric structures to concrete statistical models. A *statistical model* is a family of probability distributions parameterized by a set of parameters.

Definition 14 (Statistical Models). A *statistical model* $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ is a family of probability density functions (or probability mass functions) on a sample space $\mathcal{X}$, where $\Theta \subseteq \mathbb{R}^D$ is an open subset called the *parameter space*. We assume that the mapping $\theta \mapsto p(x;\theta)$ is smooth and that the distributions are identifiable (i.e., different parameters yield different distributions).

Given a statistical model, we can naturally construct a statistical manifold structure. The most fundamental geometric object is the *Fisher information matrix*.

Definition 15 (Fisher Information Matrix). For a statistical model $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$, the *Fisher information matrix* $I(\theta)$ at parameter θ is defined as

$$I_{ij}(\theta) = \mathrm{E}_\theta\left[\frac{\partial \log p(X;\theta)}{\partial \theta^i}\frac{\partial \log p(X;\theta)}{\partial \theta^j}\right], \tag{18}$$

where E_θ denotes expectation with respect to $p(x;\theta)$. Equivalently, using the score function $s_i(x;\theta) = \frac{\partial \log p(x;\theta)}{\partial \theta^i}$, we have

$$I_{ij}(\theta) = \mathrm{E}_\theta[s_i(X;\theta)s_j(X;\theta)]. \tag{19}$$

The Fisher information matrix provides a natural Riemannian metric on the parameter space, known as the *Fisher-Rao metric*.

Definition 16 (Fisher-Rao Metric). The *Fisher-Rao metric* g on a statistical model $\mathcal{P}$ is defined by taking the Fisher information matrix as the metric tensor:

$$g_{ij}(\theta) = I_{ij}(\theta) = \mathrm{E}_\theta \left[\frac{\partial \log p(X;\theta)}{\partial \theta^i} \frac{\partial \log p(X;\theta)}{\partial \theta^j} \right]. \tag{20}$$

This metric endows the parameter space Θ with a Riemannian manifold structure.

The Fisher-Rao metric has several important properties. It is invariant under sufficient statistics and provides a natural measure of distance between probability distributions. The distance induced by this metric is called the *Fisher-Rao distance*.

Definition 17 (Amari-Chentsov Cubic Tensor). For a statistical model $\mathcal{P}$, the *Amari-Chentsov cubic tensor* (or *skewness tensor*) is defined as

$$C_{ijk}(\theta) = \mathrm{E}_\theta \left[\frac{\partial \log p(X;\theta)}{\partial \theta^i} \frac{\partial \log p(X;\theta)}{\partial \theta^j} \frac{\partial \log p(X;\theta)}{\partial \theta^k} \right]. \tag{21}$$

This tensor measures the third-order moments of the score function and captures the asymmetry of the statistical model.

Together, the Fisher-Rao metric g and the Amari-Chentsov tensor C define a statistical manifold structure on the parameter space Θ.

1.5 Exponential Families and Mixture Families

Two important classes of statistical models that play a central role in information geometry are *exponential families* and *mixture families*. These families have particularly nice geometric properties.

Definition 18 (Exponential Families). An *exponential family* is a statistical model of the form

$$p(x;\theta) = \exp\left(\theta^i F_i(x) - \psi(\theta) + k(x)\right), \tag{22}$$

where $\theta = (\theta^1, \ldots, \theta^D) \in \Theta$ are the *natural parameters*, $F_i(x)$ are the *sufficient statistics*, $\psi(\theta)$ is the *log-partition function* (or *cumulant generating function*), and $k(x)$ is the *carrier measure*. The log-partition function ensures normalization:

$$\psi(\theta) = \log \int \exp(\theta^i F_i(x) + k(x))\, dx. \tag{23}$$

Exponential families include many common distributions such as Gaussian, Poisson, exponential, and multinomial distributions. They have the important property that the Fisher information matrix is given by the Hessian of the log-partition function.

Proposition 1. For an exponential family with natural parameters θ, the Fisher information matrix is

$$I_{ij}(\theta) = \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}. \tag{24}$$

Definition 19 (Mixture Families). A *mixture family* is a statistical model of the form

$$p(x;\eta) = \sum_{i=1}^{D+1} \eta^i p_i(x), \tag{25}$$

where $\{p_i(x)\}_{i=1}^{D+1}$ are $D+1$ linearly independent probability distributions, and $\eta = (\eta^1, \dots, \eta^D)$ are the *mixture parameters* satisfying $\eta^i \geq 0$ and $\sum_{i=1}^{D+1} \eta^i = 1$.

Mixture families include finite mixture models and are important in clustering and density estimation. They have complementary geometric properties to exponential families.

1.6 Dually Flat Manifolds

A particularly important class of statistical manifolds are those that are *dually flat*, meaning that both the $\nabla^{(1)}$-connection and the $\nabla^{(-1)}$-connection are flat (have vanishing curvature).

Definition 20 (Dually Flat Manifolds). A statistical manifold $(\mathcal{M}, g, \nabla, \nabla^*)$ is called *dually flat* if both connections ∇ and ∇^* are flat, i.e., their Riemann curvature tensors vanish identically.

Dually flat manifolds have remarkable properties. They admit two special coordinate systems: the *affine coordinates* with respect to ∇ and the *dual affine coordinates* with respect to ∇^*. These coordinate systems are related by a *Legendre transformation*.

Theorem 5. On a dually flat manifold $(\mathcal{M}, g, \nabla, \nabla^*)$, there exist two coordinate systems $\theta = (\theta^1, \dots, \theta^D)$ and $\eta = (\eta_1, \dots, \eta_D)$ such that:

1. The θ-coordinates are ∇-affine (geodesics are straight lines in these coordinates).
2. The η-coordinates are ∇^*-affine (geodesics are straight lines in these coordinates).
3. The two coordinate systems are related by a Legendre transformation:

$$\eta_i = \frac{\partial \psi(\theta)}{\partial \theta^i}, \quad \theta^i = \frac{\partial \phi(\eta)}{\partial \eta_i}, \tag{26}$$

   where ψ and ϕ are convex functions satisfying $\psi(\theta) + \phi(\eta) - \theta^i \eta_i = 0$.
4. The metric tensor in these coordinates is given by

$$g_{ij}(\theta) = \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}, \quad g^{ij}(\eta) = \frac{\partial^2 \phi(\eta)}{\partial \eta_i \partial \eta_j}. \tag{27}$$

Exponential families and mixture families are examples of dually flat manifolds. For exponential families, the natural parameters θ are the $\nabla^{(1)}$-affine coordinates, while the expectation parameters $\eta = \mathbb{E}[F_i(X)]$ are the $\nabla^{(-1)}$-affine coordinates.

1.7 Divergences and Distances

Divergences provide a way to measure the "distance" between probability distributions. While not true metrics (they may not be symmetric or satisfy the triangle inequality), divergences play a crucial role in information geometry.

Definition 21 (Divergences). A *divergence* $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$ on a statistical model $\mathcal{P}$ is a function satisfying:

1. $D(p\|q) \geq 0$ for all $p, q \in \mathcal{P}$.
2. $D(p\|q) = 0$ if and only if $p = q$.
3. For smooth p, q, the second-order Taylor expansion around $p = q$ yields the Fisher information metric.

The most fundamental divergence in information theory is the *Kullback-Leibler (KL) divergence*.

Definition 22 (Kullback-Leibler Divergence). The *Kullback-Leibler divergence* (or *relative entropy*) between two probability distributions p and q is defined as

$$D_{\mathrm{KL}}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx. \tag{28}$$

For discrete distributions, the integral is replaced by a sum.

The KL divergence is not symmetric, but it is closely related to the geometric structure of statistical manifolds. For exponential families, the KL divergence can be expressed in terms of the dual coordinate systems.

Proposition 2. For an exponential family with natural parameters θ_p, θ_q and expectation parameters η_p, η_q, the KL divergence is

$$D_{\mathrm{KL}}(p\|q) = \psi(\theta_q) - \psi(\theta_p) - \theta_q^i(\eta_{p,i} - \eta_{q,i}) = \phi(\eta_p) - \phi(\eta_q) - \eta_{p,i}(\theta_q^i - \theta_p^i). \tag{29}$$

More generally, we can define a family of α-divergences that interpolate between different divergences.

Definition 23 (α-Divergences). The *α-divergence* between two probability distributions p and q is defined as

$$D^{(\alpha)}(p\|q) = \frac{4}{1-\alpha^2} \int \left(1 - p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\right) dx, \tag{30}$$

for $\alpha \neq \pm 1$. In the limit $\alpha \to 1$, we recover the KL divergence $D_{\mathrm{KL}}(p\|q)$, and in the limit $\alpha \to -1$, we obtain $D_{\mathrm{KL}}(q\|p)$.

The α-divergences are related to the α-connections: the second-order expansion of an α-divergence yields the Fisher information metric, and the third-order term is related to the cubic tensor that defines the α-connections.

1.8 Applications and Further Directions

Information geometry has found numerous applications across various fields. We briefly mention a few key applications.

1. *Bayesian Hypothesis Testing*: The geometric structure of statistical manifolds provides insights into optimal decision rules and the geometry of hypothesis testing problems.

2. *Statistical Inference*: Maximum likelihood estimation, efficient estimators, and the Cramér-Rao bound have natural geometric interpretations in terms of the Fisher information metric.

3. *Machine Learning*: Information geometry has been applied to neural networks, clustering algorithms (e.g.,

k-means with Bregman divergences), and optimization on manifolds.

4. *Neuroscience*: The geometric structure of neural activity spaces, information coding, and neural computation can be analyzed using information-geometric methods.

5. *Quantum Information*: Quantum information geometry extends these concepts to quantum states, providing geometric insights into quantum estimation and quantum information processing.

The field continues to develop, with active research in non-parametric information geometry, infinite-dimensional statistical manifolds, and applications to modern machine learning and data science.

2 Information Geometry of Neural Coding

The application of information geometry to neural coding provides a powerful framework for understanding how populations of neurons represent and transmit information. This section explores the geometric structure of neural population codes, with particular emphasis on the role of noise correlations and their impact on information transmission [SR21, DLM$^+$].

2.1 Neural Population Codes and Noise Correlations

Neurons in the brain represent information through their collective activity patterns. The fidelity of this neural population code depends critically on the structure of variability in neural responses, particularly on whether and how variability in one neuron's response is shared with other neurons. This shared variability, known as *noise correlations*, plays a fundamental role in determining the information-carrying capacity of neural populations.

Definition 24 (Neural Population Code). A *neural population code* is a mapping from stimuli $s \in \mathcal{S}$ to probability distributions over neural response patterns $\mathbf{r} = (r_1, r_2, \ldots, r_N) \in \mathcal{R}^N$, where N is the number of neurons in the population. The code is characterized by the conditional probability distribution $p(\mathbf{r}|s)$.

The geometric approach to neural coding views the space of possible response patterns as a statistical manifold, where each stimulus corresponds to a point on this manifold. The Fisher information metric provides a natural measure of how discriminable different stimuli are based on the neural responses.

Definition 25 (Noise Correlations). For a neural population with responses $\mathbf{r} = (r_1, r_2, \ldots, r_N)$, the *noise correlations* are defined as the correlations in the variability of responses around their mean:

$$\Sigma_{ij} = \mathrm{E}[(r_i - \mathrm{E}[r_i|s])(r_j - \mathrm{E}[r_j|s])|s], \tag{31}$$

where Σ is the noise covariance matrix. The diagonal elements Σ_{ii} represent the variance of individual neurons, while off-diagonal elements capture shared variability.

The structure of noise correlations has profound implications for information transmission. Two decades of theoretical and experimental work have revealed that noise correlations can either enhance or limit the information-carrying capacity of neural populations, depending on their fine structure relative to the signal correlations.

2.2 Geometric Framework for Neural Coding

The geometric approach to neural coding leverages information geometry to understand how the structure of noise correlations affects stimulus discriminability. The key insight is that the Fisher information matrix, which determines the local geometry of the response manifold, depends on both the mean responses and the noise correlations.

Definition 26 (Fisher Information for Neural Populations). For a neural population code with conditional distribution $p(\mathbf{r}|s)$, the *Fisher information* about stimulus s is given by

$$I(s) = \mathrm{E}\left[\left(\frac{\partial \log p(\mathbf{r}|s)}{\partial s}\right)^2 \middle| s\right]. \tag{32}$$

For a multivariate Gaussian response model with mean $\boldsymbol{\mu}(s)$ and covariance $\Sigma(s)$, this becomes

$$I(s) = \left(\frac{\partial \boldsymbol{\mu}(s)}{\partial s}\right)^T \Sigma(s)^{-1} \left(\frac{\partial \boldsymbol{\mu}(s)}{\partial s}\right) + \frac{1}{2}\mathrm{Tr}\left[\Sigma(s)^{-1}\frac{\partial \Sigma(s)}{\partial s}\Sigma(s)^{-1}\frac{\partial \Sigma(s)}{\partial s}\right]. \tag{33}$$

The geometric picture emerges when we consider how noise correlations affect the Fisher information. The Fisher information matrix defines a Riemannian metric on the stimulus space, where distances between stimuli correspond to their discriminability based on neural responses.

Proposition 3 (Geometric Interpretation of Noise Correlations). The effect of noise correlations on information transmission can be understood geometrically:

1. Noise correlations that align with the signal direction (the direction of mean response changes) reduce Fisher information and limit discriminability.
2. Noise correlations orthogonal to the signal direction do not affect Fisher information.
3. The optimal noise correlation structure for information transmission depends on the geometry of the signal manifold.

This geometric framework provides a unified perspective on the diverse effects of noise correlations observed in experimental studies. Rather than asking whether correlations are "good" or "bad" for information transmission, the geometric approach reveals that the impact depends on the relationship between the correlation structure and the signal geometry.

2.3 Discriminability and the Geometry of Response Manifolds

The ability to discriminate between stimuli is fundamentally a geometric property of the response manifold. Information geometry provides tools to quantify this discriminability and understand how it depends on the structure of neural responses.

Definition 27 (Discriminability). The *discriminability* between two stimuli s_1 and s_2 based on neural responses is measured by the Fisher-Rao distance between the corresponding probability distributions $p(\mathbf{r}|s_1)$ and $p(\mathbf{r}|s_2)$ on the response manifold. For small stimulus differences, this is approximated by

$$d(s_1, s_2) \approx \sqrt{(s_1 - s_2)^T I(s)(s_1 - s_2)}, \tag{34}$$

where $I(s)$ is the Fisher information matrix evaluated at an intermediate stimulus value.

The geometric approach reveals that the most discriminable stimulus directions correspond to the eigenvectors of the Fisher information matrix with the largest eigenvalues. These directions represent the stimulus dimensions along which the neural population is most sensitive.

Theorem 6 (Optimal Stimulus Directions). For a neural population code with Fisher information matrix $I(s)$, the stimulus directions that maximize discriminability are given by the eigenvectors of $I(s)$ corresponding to the largest eigenvalues. These directions are orthogonal with respect to the Fisher information metric.

This result has important implications for understanding how neural populations encode information. It suggests that the brain may be optimized to discriminate stimuli along certain privileged directions, which correspond to the principal axes of the Fisher information matrix.

2.4 Information Geometry of Retinal Representations

The application of information geometry to retinal ganglion cell populations provides a concrete example of how geometric methods can reveal the structure of neural coding. The retina transforms visual stimuli into patterns of action potentials, and the geometry of this transformation determines the fidelity of visual information transmission.

Example 2 (Retinal Representation Manifold). Consider a population of retinal ganglion cells responding to naturalistic visual stimuli. The joint probability distribution of neural responses conditioned on the stimulus can be modeled using a stochastic encoding model. The Fisher information metric over the stimulus space reveals the most discriminable stimulus directions.

Key findings from information-geometric analysis of retinal representations include:

1. The most discriminable stimulus directions vary substantially across different stimuli, indicating that the retina adapts its coding strategy to the local stimulus structure.
2. The most discriminative response modes are often aligned with the most stochastic modes, suggesting that noise correlations are information-limiting rather than information-enhancing under natural scenes.
3. Population coding benefits from complementary coding strategies, helping to equalize the information carried by different firing rates.

The geometric analysis of retinal coding reveals that the structure of noise correlations plays a critical role in determining information transmission. Under naturalistic conditions, noise correlations often limit rather than enhance information, contrary to what might be expected from simplified models.

2.5 The Role of Correlation Structure

The fine structure of noise correlations, not just their overall magnitude, determines their impact on information transmission. Information geometry provides tools to characterize and understand this fine structure.

Definition 28 (Signal and Noise Subspaces). For a neural population code, we can decompose the response space into *signal* and *noise* subspaces:

1. The *signal subspace* is spanned by the directions along which mean responses vary with the stimulus: $\text{span}\{\partial\boldsymbol{\mu}(s)/\partial s_i\}$.
2. The *noise subspace* is the orthogonal complement with respect to the Fisher information metric.

The impact of noise correlations depends on how they project onto these subspaces.

This decomposition provides a geometric framework for understanding when noise correlations help or hinder information transmission. Correlations that project strongly onto the signal subspace reduce discriminability, while correlations confined to the noise subspace have minimal impact.

Proposition 4 (Information-Limiting Correlations). Noise correlations are *information-limiting* if they align with the signal direction, meaning that the principal eigenvector of the noise covariance matrix Σ is parallel to the direction of mean response changes $\partial \boldsymbol{\mu}/\partial s$. In this case, the Fisher information is bounded by the inverse of the noise variance along the signal direction, regardless of the number of neurons.

This result explains why certain correlation structures can fundamentally limit information transmission, even in large populations. The geometric perspective makes it clear that the problem is not simply the magnitude of correlations, but their alignment with the signal structure.

2.6 Applications to Neural Circuit Analysis

The geometric framework for neural coding has been applied to understand how circuit properties affect information transmission. In particular, the competition between membrane and synaptic response timescales shapes the information geometry of neural network responses.

Example 3 (Excitatory-Inhibitory Networks). Consider a network model of excitatory and inhibitory neural populations. The information geometry of the network's response manifold depends on the balance between excitation and inhibition, as well as the relative timescales of membrane and synaptic dynamics [CB24].

Key geometric insights include:

1. The hyperbolic embedding of the response manifold reveals the statistical parameters to which the model behavior is most sensitive.
2. The ranking of these sensitive coordinates changes with the balance of excitation and inhibition.
3. The geometry of the response manifold provides a quantitative measure of how circuit properties shape information processing.

This application demonstrates how information geometry can bridge the gap between circuit-level properties and information-theoretic measures of neural coding. The geometric structure of the response manifold encodes information about both the circuit dynamics and the resulting coding properties.

2.7 Summary and Future Directions

The geometric approach to neural coding provides a unified framework for understanding how noise correlations affect information transmission. Key insights include:

1. The impact of noise correlations depends on their geometric relationship to the signal structure, not just their magnitude.
2. The Fisher information metric provides a natural measure of stimulus discriminability that incorporates both mean responses and noise correlations.

3. The geometry of response manifolds reveals the stimulus dimensions along which neural populations are most sensitive.

4. Information-limiting correlations align with signal directions, fundamentally constraining information transmission regardless of population size.

Future directions include extending the geometric framework to non-Gaussian response models, understanding how learning shapes the geometry of neural representations, and applying information-geometric methods to analyze large-scale neural recordings.

3 Feature Learning and Representational Geometry

The study of how neural networks learn task-relevant features is fundamental to understanding both biological and artificial intelligence systems. Information geometry provides a powerful framework for analyzing the evolution of neural representations during learning, revealing how task-relevant information becomes integrated into the geometry of representational manifolds [CLWC25, MCC25].

3.1 The Lazy-Rich Dichotomy and Beyond

Recent theoretical work has categorized neural network learning into two regimes: the *rich regime*, where networks actively learn task-relevant features, and the *lazy regime*, where networks behave like random feature models with minimal feature learning. However, this simple dichotomy overlooks the rich diversity of feature learning strategies that emerge from different learning algorithms, network architectures, and data properties.

Definition 29 (Lazy and Rich Learning Regimes). Consider a neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta \in \Theta$:

1. In the *lazy regime*, the network parameters θ remain close to their initialization θ_0, and the network behaves approximately as a linear model: $f_\theta(x) \approx f_{\theta_0}(x) + \nabla_\theta f_{\theta_0}(x)^T(\theta - \theta_0)$.
2. In the *rich regime*, the network parameters undergo significant changes during training, and the learned features $\phi_\theta : \mathcal{X} \to \mathcal{H}$ (where $\mathcal{H}$ is a hidden representation space) adapt to the task structure.

The lazy-rich dichotomy provides a useful starting point, but real learning dynamics often exhibit more nuanced behavior. Networks may transition between regimes, or exhibit rich learning in some layers while remaining lazy in others. A more comprehensive understanding requires analyzing the geometry of learned representations.

3.2 Representational Manifolds and Their Geometry

The key insight of the geometric approach to feature learning is to characterize how task-relevant information shapes the geometry of representational manifolds, rather than inspecting individual learned features. This manifold-centric perspective reveals distinct learning stages and strategies.

Definition 30 (Representational Manifold). For a neural network with hidden representation $\mathbf{h} = \phi_\theta(\mathbf{x}) \in \mathbb{R}^d$, the *representational manifold* $\mathcal{M}_\theta$ is the set of all possible hidden representations:

$$\mathcal{M}_\theta = \{\phi_\theta(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\} \subseteq \mathbb{R}^d. \tag{35}$$

The geometry of this manifold encodes information about how the network represents different inputs and their relationships.

The geometry of the representational manifold evolves during learning, and these geometric changes reveal how task-relevant information becomes integrated into the representation. Information geometry provides tools to quantify these changes.

Definition 31 (Manifold Geometry Metrics). For a representational manifold $\mathcal{M}_\theta$, we can define several geometric quantities:

1. The *intrinsic dimension* of the manifold, which measures the effective dimensionality of the representation.
2. The *curvature* of the manifold, which captures how the representation space is curved.
3. The *separation* between manifolds corresponding to different classes, which determines classification performance.
4. The *alignment* of the manifold with task-relevant directions, which measures how well the representation supports the task.

These geometric metrics provide a quantitative framework for understanding feature learning beyond the binary lazy-rich classification.

3.3 Manifold Untangling During Learning

A central observation is that as networks learn task-relevant features, the geometry of representational manifolds undergoes a process of *untangling*, where manifolds corresponding to different classes become more separated and better aligned with task-relevant directions.

Definition 32 (Manifold Untangling). Consider a classification task with classes $\{1, 2, \ldots, C\}$, where each class c corresponds to a set of inputs $\mathcal{X}_c \subseteq \mathcal{X}$. The representational manifolds $\mathcal{M}_{\theta,c} = \{\phi_\theta(\mathbf{x}) : \mathbf{x} \in \mathcal{X}_c\}$ are said to *untangle* during learning if:

1. The separation between manifolds increases: $\min_{\mathbf{h}_i \in \mathcal{M}_{\theta,i}, \mathbf{h}_j \in \mathcal{M}_{\theta,j}} \|\mathbf{h}_i - \mathbf{h}_j\|$ increases for $i \neq j$.
2. The manifolds become more linearly separable, meaning there exists a hyperplane that separates the manifolds with increasing margin.
3. The intrinsic dimensionality of task-irrelevant variations decreases relative to task-relevant variations.

The process of manifold untangling can be quantified using information-geometric measures. The Fisher information metric on the representational manifold provides a natural way to measure how the geometry changes during learning.

Proposition 5 (Geometric Evolution During Learning). As a neural network learns, the representational manifold $\mathcal{M}_\theta$ evolves such that:

1. The Fisher information metric becomes more aligned with task-relevant directions.

2. The curvature of the manifold decreases in task-relevant directions, making the manifold more "flat" along these directions.

3. The volume of the manifold in task-irrelevant directions contracts, while expanding in task-relevant directions.

This geometric perspective reveals that feature learning is fundamentally about reshaping the geometry of the representational space to support the task.

3.4 Learning Stages and Geometric Transitions

The geometric analysis of representational manifolds reveals that learning often proceeds through distinct stages, each characterized by different geometric properties. These stages are not captured by the simple lazy-rich dichotomy.

Definition 33 (Learning Stages from Geometry). Based on the evolution of manifold geometry, we can identify several learning stages:

1. *Initialization stage*: The manifold geometry is determined by random initialization, with high curvature and poor alignment with task structure.
2. *Feature discovery stage*: The manifold begins to untangle, with task-relevant directions emerging and curvature decreasing along these directions.
3. *Feature refinement stage*: The manifold geometry stabilizes, with fine-tuning of the alignment between the representation and task structure.
4. *Specialization stage*: The manifold becomes highly specialized to the training data, potentially leading to overfitting if not regularized.

These stages can be identified by monitoring geometric metrics such as manifold separation, curvature, and alignment throughout training. The transitions between stages often correspond to changes in the learning dynamics, such as shifts in the effective learning rate or the emergence of new features.

Theorem 7 (Geometric Characterization of Learning Stages). For a neural network learning a classification task, the learning stages can be characterized by the evolution of the following geometric quantities:

1. The *manifold capacity*, which measures the maximum number of linearly separable manifolds that can be embedded in the representation space.
2. The *manifold dimension*, which measures the effective dimensionality of each class manifold.
3. The *manifold radius*, which measures the spread of each class manifold.

Transitions between learning stages correspond to qualitative changes in how these quantities evolve.

This geometric characterization provides a more nuanced view of learning than the lazy-rich dichotomy, revealing the rich structure of feature learning dynamics.

3.5 Context-Dependent Nonlinearity and Manifold Geometry

An important extension of the geometric framework is to account for context-dependent nonlinearity, where the geometry of representational manifolds depends on contextual information. This is particularly relevant for understanding how neural systems process information in context-dependent ways.

Definition 34 (Context-Dependent Manifold Geometry). Consider a neural network that processes inputs $\mathbf{x}$ in the context of additional information $\mathbf{c}$. The representational manifold becomes context-dependent:

$$\mathcal{M}_{\theta,\mathbf{c}} = \{\phi_\theta(\mathbf{x}, \mathbf{c}) : \mathbf{x} \in \mathcal{X}\}. \tag{36}$$

The geometry of this manifold depends on the context $\mathbf{c}$, and context-dependent nonlinearity refers to how this geometry changes with context.

Context-dependent nonlinearity is ubiquitous in neural systems, where the same stimulus can be processed differently depending on context. The geometric framework provides tools to quantify and understand these context-dependent effects.

Proposition 6 (Context-Dependent Manifold Capacity). The *context-dependent manifold capacity* depends on both the geometry of the manifolds and the correlations between contexts. For manifolds with context-dependent geometry, the capacity can be expressed as a function of:

1. The intrinsic geometry of each context-specific manifold.
2. The correlations between different contexts.
3. The alignment between context-dependent and context-independent features.

This framework allows for a more expressive analysis of neural representations, capturing how context shapes the geometry of representational manifolds and enabling the study of context-dependent computation.

3.6 Structural Inductive Biases and Generalization

The geometric analysis of feature learning also sheds light on structural inductive biases, which are the architectural and algorithmic properties that shape how networks learn features. These biases are reflected in the geometry of learned representations.

Definition 35 (Structural Inductive Bias). A *structural inductive bias* is a property of the learning algorithm or network architecture that influences which features are learned and how they are organized. Examples include:

1. *Architectural biases*: Properties of the network architecture (e.g., convolutional structure, attention mechanisms) that favor certain types of features.
2. *Algorithmic biases*: Properties of the learning algorithm (e.g., gradient descent, regularization) that shape the learning dynamics.
3. *Data biases*: Properties of the training data that influence which features are most useful.

The geometry of learned representations encodes these biases. For example, convolutional architectures lead to representations with translation-invariant geometry, while attention mechanisms create representations with geometry that adapts to input structure.

Proposition 7 (Geometry and Generalization). The geometry of learned representations is predictive of generalization performance:

1. Representations with lower intrinsic dimensionality in task-irrelevant directions generalize better.
2. Representations with geometry that is robust to input perturbations generalize better to out-of-distribution data.
3. The alignment between the representation geometry and the geometry of the data distribution determines generalization.

This connection between geometry and generalization provides a principled way to understand and improve generalization in neural networks.

3.7 Applications to Neuroscience and Machine Learning

The geometric framework for feature learning has applications in both neuroscience and machine learning. In neuroscience, it provides tools to understand how biological neural circuits learn and represent information. In machine learning, it offers insights into how artificial networks learn and how to design better architectures and training procedures.

Example 4 (Neural Circuit Analysis). The geometric analysis of neural representations can reveal how biological neural circuits implement feature learning. By analyzing the geometry of neural activity manifolds, we can:

1. Identify which features are learned at different stages of development or learning.
2. Understand how circuit properties (e.g., connectivity patterns, plasticity rules) shape the geometry of representations.
3. Predict how changes in circuit properties will affect learning and representation.

Example 5 (Deep Network Analysis). The geometric framework provides tools to analyze and improve deep neural networks:

1. Monitor the geometry of representations during training to understand learning dynamics.
2. Design architectures that promote desirable geometric properties (e.g., better manifold separation).
3. Develop training procedures that explicitly optimize geometric metrics related to generalization.

3.8 Summary and Future Directions

The geometric approach to feature learning provides a comprehensive framework for understanding how neural networks learn task-relevant features. Key insights include:

1. Feature learning can be understood as the evolution of representational manifold geometry.
2. The lazy-rich dichotomy is insufficient; learning proceeds through multiple stages with distinct geometric properties.
3. Manifold untangling, the process by which class manifolds become separated and aligned with task structure, is a central mechanism of feature learning.
4. Context-dependent nonlinearity can be understood through context-dependent manifold geometry.
5. The geometry of learned representations encodes structural inductive biases and is predictive of generalization.

Future directions include developing more sophisticated geometric metrics, understanding how different architectures and training procedures shape manifold geometry, and applying the framework to analyze large-scale neural recordings and deep networks.

4 Gradient Learning and Natural Gradient Methods

The application of information geometry to optimization and learning algorithms has led to the development of natural gradient methods, which account for the geometric structure of parameter spaces. These methods provide principled approaches to gradient-based learning that respect the intrinsic geometry of statistical manifolds [Ama83, Nie20].

4.1 Natural Gradient Descent

Standard gradient descent methods treat the parameter space as Euclidean, ignoring the geometric structure induced by the statistical model. Natural gradient descent, introduced by Amari, incorporates the Fisher information metric to define gradients that respect the intrinsic geometry of the parameter space.

Definition 36 (Natural Gradient). For a loss function $\mathcal{L}(\theta)$ defined on a statistical manifold with Fisher information metric $I(\theta)$, the *natural gradient* $\tilde{\nabla}_\theta \mathcal{L}$ is defined as

$$\tilde{\nabla}_\theta \mathcal{L} = I(\theta)^{-1} \nabla_\theta \mathcal{L}, \tag{37}$$

where $\nabla_\theta \mathcal{L}$ is the standard (Euclidean) gradient. The natural gradient is the steepest descent direction with respect to the Fisher-Rao metric, rather than the Euclidean metric.

The natural gradient accounts for the fact that distances in parameter space should be measured using the Fisher information metric, which reflects how changes in parameters affect the probability distribution. This leads to more efficient optimization, particularly in high-dimensional parameter spaces with strong curvature.

Proposition 8 (Optimality of Natural Gradient). The natural gradient direction is the direction of steepest descent on the statistical manifold, meaning it minimizes the loss function most efficiently per unit distance traveled on the manifold (as measured by the Fisher-Rao metric).

This optimality property makes natural gradient methods particularly effective for learning in overparameterized models, where the parameter space has complex geometry.

4.2 Information Geometry of Optimization

The geometric perspective on optimization reveals that the efficiency of learning algorithms depends on how well they account for the geometry of the loss landscape. Information geometry provides tools to characterize this geometry and design algorithms that exploit it.

Definition 37 (Loss Landscape Geometry). For a statistical model with parameters $\theta \in \Theta$ and loss function $\mathcal{L}(\theta)$, the *loss landscape* is the graph of $\mathcal{L}$ over the parameter space. The geometry of this landscape is characterized by:

1. The *Hessian matrix* $H_{ij} = \partial^2 \mathcal{L} / \partial\theta^i \partial\theta^j$, which measures local curvature.
2. The *Fisher information matrix* $I_{ij}(\theta)$, which measures the geometry of the statistical manifold.
3. The relationship between H and I, which determines how well standard gradient methods approximate natural gradient methods.

When the loss function is the negative log-likelihood, there is a close relationship between the Hessian and the

Fisher information matrix. This relationship is the foundation for many approximate natural gradient methods.

Theorem 8 (Relationship Between Hessian and Fisher Information). For a loss function $\mathcal{L}(\theta) = -\log p(\mathbf{x}|\theta)$ (negative log-likelihood), the expected Hessian equals the Fisher information matrix:

$$\mathbb{E}_{\mathbf{x}\sim p(\cdot|\theta)}[H(\theta)] = I(\theta). \tag{38}$$

This relationship allows the Fisher information matrix to serve as an approximation to the Hessian, enabling efficient second-order optimization methods.

This theorem connects information geometry to optimization theory, showing that the Fisher information matrix provides a natural preconditioner for gradient-based optimization.

4.3 Approximate Natural Gradient Methods

Computing the exact natural gradient requires inverting the Fisher information matrix, which is computationally expensive for large parameter spaces. Several approximation methods have been developed to make natural gradient methods practical.

Definition 38 (Kronecker-Factored Approximate Curvature (K-FAC)). For neural networks with layer-wise structure, the Fisher information matrix can be approximated using a Kronecker factorization:

$$I(\theta) \approx \bigotimes_{l=1}^{L} A_l \otimes B_l, \tag{39}$$

where A_l and B_l are smaller matrices associated with layer l, and $\otimes$ denotes the Kronecker product. This factorization allows efficient inversion and matrix-vector products.

K-FAC and related methods make natural gradient descent practical for large neural networks by exploiting the structure of the Fisher information matrix. These methods have been shown to accelerate training and improve generalization.

Proposition 9 (Efficiency of K-FAC). The Kronecker factorization reduces the computational complexity of natural gradient computation from $O(d^3)$ to $O(d^{3/2})$ for a network with d parameters, while maintaining most of the benefits of exact natural gradient descent.

Other approximation methods include diagonal approximations, block-diagonal approximations, and methods that use low-rank approximations to the Fisher information matrix.

4.4 Gradient Learning on Manifolds

The geometric perspective extends beyond natural gradient methods to general gradient learning on manifolds. When the parameter space itself is a manifold (not just a statistical manifold), gradient-based learning must account for the manifold structure.

Definition 39 (Gradient on a Manifold). For a function $f : \mathcal{M} \to \mathbb{R}$ defined on a Riemannian manifold $(\mathcal{M}, g)$, the *gradient* $\nabla_g f$ is the unique vector field such that for any vector field X,

$$g(\nabla_g f, X) = X(f) = df(X), \tag{40}$$

where df is the differential of f. In local coordinates, this becomes

$$(\nabla_g f)^i = g^{ij} \frac{\partial f}{\partial x^j}. \tag{41}$$

Gradient descent on manifolds requires projecting the gradient back onto the manifold and updating parameters along geodesics (or approximations thereof). This is the foundation for optimization on manifolds such as the Stiefel manifold, the Grassmannian, and other matrix manifolds.

Proposition 10 (Geodesic Gradient Descent). Gradient descent on a Riemannian manifold $(\mathcal{M}, g)$ updates parameters along geodesics:

$$\theta_{t+1} = \exp_{\theta_t}(-\eta \nabla_g \mathcal{L}(\theta_t)), \tag{42}$$

where $\exp_p(v)$ is the exponential map at point p in direction v, and η is the learning rate. This ensures that updates remain on the manifold.

This geometric approach to optimization is particularly important when parameters have constraints (e.g., orthogonality constraints, unit norm constraints) that define a manifold structure.

4.5 Information Geometry and Generalization

The geometry of the parameter space, as measured by the Fisher information metric, is closely related to generalization. Flat regions of the parameter space (where the Fisher information is small) tend to generalize better than sharp regions.

Definition 40 (Flatness and Generalization). A region of parameter space is *flat* if the Fisher information matrix has small eigenvalues, meaning that small changes in parameters lead to small changes in the probability distribution. Flat minima are associated with better generalization because they are robust to parameter perturbations.

The connection between flatness and generalization provides a geometric perspective on why certain optimization methods (e.g., those that find flat minima) generalize better than others.

Proposition 11 (Flat Minima and Generalization). Minima with smaller Fisher information (flatter minima) tend to generalize better because:

1. They are more robust to parameter perturbations, reducing sensitivity to training set variations.

2. They correspond to simpler models in the sense of minimum description length.

3. They have better PAC-Bayes generalization bounds that depend on the Fisher information.

This geometric perspective on generalization motivates optimization methods that explicitly seek flat minima, such as sharpness-aware minimization and methods that regularize the Fisher information.

4.6 Adaptive Learning Rates and Geometry

The Fisher information matrix also provides a principled way to set adaptive learning rates. Methods that use the Fisher information as a preconditioner automatically adapt the learning rate to the local geometry of the parameter space.

Definition 41 (Adaptive Learning Rate from Fisher Information). Natural gradient descent implicitly uses an adaptive learning rate:

$$\theta_{t+1} = \theta_t - \eta I(\theta_t)^{-1} \nabla_\theta \mathcal{L}(\theta_t), \tag{43}$$

where the effective learning rate in direction i is η/λ_i, with λ_i being the i-th eigenvalue of $I(\theta_t)$. Directions with large Fisher information (high sensitivity) get smaller effective learning rates, while directions with small Fisher information get larger effective learning rates.

This adaptive behavior is particularly important in high-dimensional parameter spaces where different directions have very different sensitivities. Standard gradient descent with a fixed learning rate can be inefficient in such settings.

Proposition 12 (Efficiency of Adaptive Learning Rates). Adaptive learning rates based on Fisher information can accelerate convergence by:

1. Taking larger steps in directions with low sensitivity (flat directions).
2. Taking smaller steps in directions with high sensitivity (steep directions).
3. Automatically adapting to the local geometry without manual tuning.

This geometric perspective explains why methods like Adam, which use adaptive learning rates, often outperform standard gradient descent, even though they don't explicitly use the Fisher information matrix.

4.7 Applications and Extensions

Natural gradient methods and geometric optimization have found applications in many areas of machine learning, including neural network training, reinforcement learning, and variational inference.

Example 6 (Neural Network Training). Natural gradient methods have been successfully applied to training deep neural networks:

1. K-FAC and related methods accelerate training by better accounting for parameter space geometry.
2. These methods often find solutions with better generalization properties.
3. They are particularly effective for training recurrent neural networks and other architectures with complex parameter dependencies.

Example 7 (Reinforcement Learning). Natural gradient methods are widely used in policy gradient reinforcement learning:

1. Natural policy gradient methods use the Fisher information of the policy distribution to improve policy updates.
2. Trust region policy optimization (TRPO) and proximal policy optimization (PPO) use approximations to natural gradients.
3. These methods provide more stable and efficient policy learning.

4.8 Summary and Future Directions

The geometric approach to gradient learning provides a principled framework for optimization that respects the intrinsic geometry of parameter spaces. Key insights include:

1. Natural gradient descent accounts for the Fisher-Rao geometry of parameter spaces, leading to more efficient optimization.
2. The Fisher information matrix provides a natural connection between information geometry and optimization theory.
3. Approximate methods like K-FAC make natural gradient methods practical for large-scale problems.
4. The geometry of parameter spaces is closely related to generalization, with flat minima generalizing better.
5. Adaptive learning rates based on Fisher information automatically adapt to local geometry.

Future directions include developing more efficient approximations to natural gradients, understanding how geometry affects generalization in deep learning, and extending geometric optimization methods to new problem domains.

5 Information Geometry of Spiking Neural Networks

Spiking neural networks (SNNs) provide a biologically plausible model of neural computation that operates on discrete spike events rather than continuous activations. The application of information geometry to SNNs reveals how the discrete, event-based nature of spiking affects the geometry of neural representations and information processing [CB24, SHW^{+}24].

5.1 Spiking Neural Networks and Event-Based Computation

Spiking neural networks differ fundamentally from traditional artificial neural networks in their use of discrete spike events to represent and transmit information. This event-based computation has important implications for the geometry of neural representations.

Definition 42 (Spiking Neural Network). A *spiking neural network* consists of neurons that generate discrete *spike events* in response to inputs. The state of neuron i at time t is characterized by:

1. The *membrane potential* $V_i(t)$, which evolves according to differential equations.
2. A *spike train* $s_i(t) = \sum_k \delta(t - t_i^k)$, where t_i^k are the spike times and δ is the Dirac delta function.
3. The relationship $s_i(t) = 1$ when $V_i(t)$ crosses a threshold θ, and $s_i(t) = 0$ otherwise.

The discrete nature of spikes creates a fundamentally different geometry compared to continuous neural networks. Information is encoded in the timing and patterns of spikes, leading to a geometry that reflects temporal structure and event-based computation.

Definition 43 (Spike Train Manifold). For a population of N spiking neurons, the *spike train manifold* is the space of all possible spike patterns over a time window $[0, T]$. Each point on this manifold corresponds to a particular pattern of spikes, and the geometry of this manifold encodes how different spike patterns relate to stimuli or computations.

The geometry of spike train manifolds is complex because spike trains are high-dimensional, sparse, and have a natural metric structure based on spike timing.

5.2 Information Geometry of Spike Patterns

The application of information geometry to spiking neural networks requires defining probability distributions over spike patterns and characterizing their geometric structure. This reveals how information is encoded in spike timing and patterns.

Definition 44 (Spike Pattern Distribution). For a population of spiking neurons responding to a stimulus s, the *spike pattern distribution* $p(\mathbf{S}|s)$ gives the probability of observing a particular spike pattern $\mathbf{S} = (s_1, s_2, \ldots, s_N)$ over a time window, where each s_i is a spike train.

The Fisher information metric on the space of spike pattern distributions provides a measure of how discriminable different stimuli are based on spike patterns. This metric depends on both the mean spike rates and the correlations in spike timing.

Proposition 13 (Fisher Information for Spike Patterns). The Fisher information about stimulus s based on spike patterns is given by

$$I(s) = \mathbf{E}\left[\left(\frac{\partial \log p(\mathbf{S}|s)}{\partial s}\right)^2 \middle| s\right], \tag{44}$$

where the expectation is over spike patterns $\mathbf{S}$. For Poisson-like spike statistics, this depends on both the sensitivity of mean firing rates to the stimulus and the structure of spike correlations.

The geometry of spike pattern distributions reveals how the temporal structure of spikes contributes to information encoding, beyond what can be captured by mean firing rates alone.

5.3 Network Dynamics and Information Geometry

The dynamics of spiking neural networks, including the competition between membrane and synaptic timescales, shape the information geometry of network responses. Understanding this relationship provides insights into how network properties affect information processing.

Definition 45 (Timescale Competition). Spiking neural networks exhibit competition between:

1. *Membrane timescales* τ_m, which determine how quickly membrane potentials respond to inputs.
2. *Synaptic timescales* τ_s, which determine how quickly synaptic inputs decay.

The ratio τ_s/τ_m determines the temporal filtering properties of the network and affects the geometry of neural responses.

The balance between excitation and inhibition, combined with timescale competition, shapes the information geometry of network responses. This geometry can be analyzed using the Fisher information metric and related geometric quantities.

Proposition 14 (Geometry of Network Responses). The information geometry of spiking network responses depends on:

1. The balance between excitation and inhibition, which affects the dimensionality of the response manifold.
2. The relative timescales of membrane and synaptic dynamics, which determine the temporal structure of responses.
3. The connectivity structure, which shapes correlations in spike patterns.

Changes in these properties lead to qualitative changes in the geometry of the response manifold.

This geometric perspective provides a quantitative framework for understanding how circuit properties affect information processing in spiking networks.

5.4 Hyperbolic Embeddings and Network Geometry

The information geometry of spiking neural networks often exhibits hyperbolic structure, where the response manifold has negative curvature. This hyperbolic geometry has important implications for how information is represented and processed.

Definition 46 (Hyperbolic Geometry of Responses). A spiking network's response manifold may have *hyperbolic geometry* if the Fisher information metric induces negative curvature. In this case, the manifold can be embedded in hyperbolic space, where distances grow exponentially rather than linearly.

Hyperbolic embeddings reveal the hierarchical structure of neural responses and identify the statistical parameters to which the network is most sensitive. This sensitivity ranking can change with network properties such as the balance of excitation and inhibition.

Theorem 9 (Hyperbolic Embedding of Response Manifolds). For spiking neural networks with certain connectivity patterns and dynamics, the response manifold can be isometrically embedded in hyperbolic space. This embedding reveals:

1. The hierarchical organization of neural responses.
2. The parameters to which the network is most sensitive.
3. How sensitivity rankings change with network properties.

This geometric characterization provides insights into how spiking networks organize and process information, revealing structure that is not apparent from standard analyses.

5.5 Spiking Networks on Riemannian Manifolds

An important extension is to consider spiking neural networks that operate on data that naturally lives on Riemannian manifolds, such as graphs or other non-Euclidean structures. This requires generalizing spiking neuron models to manifold-valued computations.

Definition 47 (Manifold-Valued Spiking Neuron). A *manifold-valued spiking neuron* operates on a Riemannian manifold $(\mathcal{M}, g)$ and generates spikes based on geodesic distances and manifold geometry. The membrane potential evolves on the manifold, and spikes occur when certain geometric conditions are met.

This generalization allows spiking networks to process structured data (e.g., graphs, manifolds) while maintaining the energy efficiency and biological plausibility of spiking computation.

Proposition 15 (Geodesic-Based Spiking). For spiking neurons on a Riemannian manifold, spike generation can be based on geodesic distances:

1. The membrane potential evolves along geodesics on the manifold.

2. Spikes occur when the geodesic distance from a reference point exceeds a threshold.

3. Information is encoded in the geometry of spike patterns on the manifold.

This geometric approach to spiking networks enables efficient processing of non-Euclidean data while preserving the advantages of event-based computation.

5.6 Learning in Spiking Networks

Training spiking neural networks presents unique challenges due to the non-differentiable nature of spike generation. Information geometry provides tools to develop gradient-based learning methods that account for the geometry of spike patterns.

Definition 48 (Surrogate Gradient Learning). *Surrogate gradient learning* approximates the gradient of the spike generation function using a smooth surrogate. This allows backpropagation through time (BPTT) to be applied to spiking networks, but requires careful handling of the temporal dynamics.

Information geometry suggests alternative approaches that respect the geometry of spike patterns. Instead of using surrogate gradients, we can define gradients directly on the manifold of spike patterns using the Fisher information metric.

Proposition 16 (Natural Gradient for Spiking Networks). Natural gradient methods for spiking networks use the Fisher information metric on spike pattern distributions:

$$\tilde{\nabla}_\theta \mathcal{L} = I(\theta)^{-1} \nabla_\theta \mathcal{L}, \tag{45}$$

where $I(\theta)$ is the Fisher information matrix for spike patterns. This accounts for the geometry of spike pattern space and can lead to more efficient learning.

This geometric approach to learning in spiking networks avoids the need for surrogate gradients and respects the intrinsic geometry of spike-based computation.

5.7 Energy Efficiency and Geometric Structure

One of the key advantages of spiking neural networks is their energy efficiency, which comes from event-based computation. The geometric structure of spike patterns is related to this energy efficiency.

Definition 49 (Energy Efficiency of Spike Patterns). The energy efficiency of a spiking network can be quantified by:

1. The *spike rate*, which determines the energy consumption.
2. The *information per spike*, which measures how much information is encoded per energy unit.
3. The *geometric efficiency*, which relates the geometry of spike patterns to information encoding efficiency.

The geometry of spike patterns affects energy efficiency because sparse, well-structured spike patterns can encode more information per spike than dense, unstructured patterns.

Proposition 17 (Geometry and Energy Efficiency). Spiking networks with geometric structure in their spike patterns (e.g., low-dimensional manifolds, hierarchical organization) tend to be more energy efficient because:

1. They can encode information using fewer spikes.

2. The geometric structure allows efficient decoding.

3. The manifold structure enables efficient routing and processing.

This connection between geometry and energy efficiency provides a principled way to design energy-efficient spiking networks.

5.8 Applications and Future Directions

Information geometry of spiking neural networks has applications in understanding biological neural circuits, designing efficient neuromorphic computing systems, and developing new learning algorithms for event-based computation.

Example 8 (Biological Neural Circuits). The geometric analysis of spiking network responses can reveal:

1. How circuit properties (excitation-inhibition balance, timescales) shape information processing.
2. The geometric structure of neural population codes.
3. How learning modifies the geometry of neural representations.

Example 9 (Neuromorphic Computing). Geometric methods can guide the design of neuromorphic systems:

1. Optimizing network architecture for energy efficiency.
2. Designing learning algorithms that respect spike-based computation.
3. Understanding the trade-offs between accuracy and energy consumption.

5.9 Summary

The application of information geometry to spiking neural networks reveals how the discrete, event-based nature of spikes creates unique geometric structure. Key insights include:

1. Spike patterns form manifolds with geometry that reflects temporal structure and event-based computation.
2. Network dynamics and timescale competition shape the information geometry of responses.
3. Hyperbolic embeddings reveal hierarchical structure and parameter sensitivity.
4. Manifold-valued spiking neurons enable processing of non-Euclidean data.
5. Natural gradient methods provide geometric approaches to learning in spiking networks.
6. The geometry of spike patterns is related to energy efficiency.

Future directions include developing more sophisticated geometric models of spike patterns, understanding how geometry emerges from network dynamics, and applying geometric methods to design more efficient and capable spiking networks.

A Proofs of Theorems

This appendix contains proofs of the main theorems stated throughout these notes. The proofs are organized by topic, following the structure of the main text.

A.1 Foundations of Information Geometry

Proof of Levi-Civita Connection Theorem. We prove the existence and uniqueness of the Levi-Civita connection. The key idea is to use the metric compatibility and torsion-free conditions to uniquely determine the Christoffel symbols.

For a metric-compatible connection, we have:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}, \tag{46}$$

where $\Gamma_{ki,j} = g_{jl}\Gamma^l_{ki}$ are the Christoffel symbols of the first kind.

For a torsion-free connection, we have:

$$\Gamma^k_{ij} = \Gamma^k_{ji}. \tag{47}$$

By cyclically permuting indices in the metric compatibility condition, we obtain:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}, \tag{48}$$

$$\partial_i g_{jk} = \Gamma_{ij,k} + \Gamma_{ik,j}, \tag{49}$$

$$\partial_j g_{ki} = \Gamma_{jk,i} + \Gamma_{ji,k}. \tag{50}$$

Adding the first two equations and subtracting the third, and using the symmetry $\Gamma_{ij,k} = \Gamma_{ji,k}$, we obtain:

$$\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki} = 2\Gamma_{ij,k}. \tag{51}$$

Therefore,

$$\Gamma_{ij,k} = \frac{1}{2}(\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki}), \tag{52}$$

and raising the index using the inverse metric:

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}). \tag{53}$$

This formula uniquely determines the connection, proving both existence and uniqueness. □

Proof of Existence of Conjugate Connection. Given a connection ∇ on $(\mathcal{M}, g)$, we need to show there exists a unique connection ∇^* such that for all vector fields X, Y, Z:

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z). \tag{54}$$

In local coordinates, this becomes:

$$\frac{\partial g_{ij}}{\partial x^k} Y^i Z^j = g_{jl} \Gamma^l_{ki} Y^i Z^j + g_{il} (\Gamma^*)^l_{kj} Y^i Z^j. \tag{55}$$

Since this must hold for all Y and Z, we have:

$$\frac{\partial g_{ij}}{\partial x^k} = g_{jl} \Gamma^l_{ki} + g_{il} (\Gamma^*)^l_{kj}. \tag{56}$$

Multiplying by g^{im} and g^{jn} and using $g^{im} g_{il} = \delta^m_l$, we obtain:

$$g^{im} g^{jn} \frac{\partial g_{ij}}{\partial x^k} = g^{jn} \Gamma^n_{ki} + g^{im} (\Gamma^*)^m_{kj}. \tag{57}$$

Rearranging and using the symmetry of the metric:

$$(\Gamma^*)^m_{kj} = g^{im} \frac{\partial g_{ij}}{\partial x^k} - g^{im} g_{jl} \Gamma^l_{ki} = g^{im} \frac{\partial g_{ij}}{\partial x^k} - \Gamma^m_{kj}. \tag{58}$$

This uniquely determines the Christoffel symbols of ∇^*, proving existence and uniqueness. □

Proof that Mean Connection Equals Levi-Civita Connection. Let $\bar{\nabla} = (\nabla + \nabla^*)/2$ be the mean connection. We need to show that $\bar{\nabla}$ is self-conjugate and equals the Levi-Civita connection.

First, we show self-conjugacy. For the mean connection:

$$\bar{\Gamma}_{ij,k} = \frac{1}{2}(\Gamma_{ij,k} + \Gamma^*_{ij,k}). \tag{59}$$

The conjugate of the mean connection has Christoffel symbols:

$$(\bar{\Gamma}^*)_{ij,k} = \frac{1}{2}((\Gamma^*)_{ij,k} + (\Gamma^*)^*_{ij,k}) = \frac{1}{2}(\Gamma^*_{ij,k} + \Gamma_{ij,k}) = \bar{\Gamma}_{ij,k}, \tag{60}$$

where we used that $(\nabla^*)^* = \nabla$. Therefore, $\bar{\nabla}$ is self-conjugate.

Now we show it equals the Levi-Civita connection. From the definition of conjugate connections:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}. \tag{61}$$

Similarly, by symmetry:

$$\partial_k g_{ij} = \Gamma^*_{ki,j} + \Gamma_{kj,i}. \tag{62}$$

Adding these two equations:

$$2\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i} + \Gamma^*_{ki,j} + \Gamma_{kj,i} = 2(\bar{\Gamma}_{ki,j} + \bar{\Gamma}_{kj,i}). \tag{63}$$

Therefore:

$$\partial_k g_{ij} = \bar{\Gamma}_{ki,j} + \bar{\Gamma}_{kj,i}, \tag{64}$$

which is exactly the metric compatibility condition. Since $\bar{\nabla}$ is also torsion-free (as the average of two torsion-free connections), it must be the Levi-Civita connection. □

Proof of Fundamental Theorem of Information Geometry. Given a statistical manifold $(\mathcal{M}, g, C)$ with cubic tensor C, we construct the α-connections as follows.

Define the Christoffel symbols of the first kind for ∇^α by:

$$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{\alpha}{2} C_{ijk}, \tag{65}$$

where $\Gamma^{(0)}_{ij,k}$ are the Christoffel symbols of the Levi-Civita connection.

Raising indices, we get:

$$\Gamma^{(\alpha),k}_{ij} = \Gamma^{(0),k}_{ij} - \frac{\alpha}{2} g^{kl} C_{ijl}. \tag{66}$$

We verify that $\nabla^{-\alpha}$ and ∇^α are conjugate. For conjugate connections, we need:

$$\partial_k g_{ij} = \Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i}. \tag{67}$$

Substituting our definitions:

$$\Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i} = \left(\Gamma^{(0)}_{ki,j} - \frac{\alpha}{2} C_{kij}\right) + \left(\Gamma^{(0)}_{kj,i} + \frac{\alpha}{2} C_{kji}\right) \tag{68}$$

$$= \Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i} - \frac{\alpha}{2}(C_{kij} - C_{kji}). \tag{69}$$

Since C is totally symmetric, $C_{kij} = C_{kji}$, so the last term vanishes. Since $\nabla^{(0)}$ is the Levi-Civita connection, we have:

$$\Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i} = \partial_k g_{ij}, \tag{70}$$

which proves the conjugacy.

For $\alpha = 0$, we have $\Gamma^{(0)}_{ij,k} = \Gamma^{(0)}_{ij,k}$, so ∇^0 is the Levi-Civita connection.

Finally, we verify the relationship with the cubic tensor:

$$\Gamma^{(-\alpha)}_{ij,k} - \Gamma^{(\alpha)}_{ij,k} = \left(\Gamma^{(0)}_{ij,k} + \frac{\alpha}{2} C_{ijk}\right) - \left(\Gamma^{(0)}_{ij,k} - \frac{\alpha}{2} C_{ijk}\right) \tag{71}$$

$$= \alpha C_{ijk}. \tag{72}$$

For $\alpha = 1$, this gives $C_{ijk} = \Gamma^{(-1)}_{ij,k} - \Gamma^{(1)}_{ij,k}$, as required. □

Proof of Fisher Information for Exponential Families. For an exponential family with natural parameters θ:

$$p(x;\theta) = \exp(\theta^i F_i(x) - \psi(\theta) + k(x)), \tag{73}$$

the log-likelihood is:

$$\log p(x;\theta) = \theta^i F_i(x) - \psi(\theta) + k(x). \tag{74}$$

The score function is:

$$\frac{\partial \log p(x;\theta)}{\partial \theta^i} = F_i(x) - \frac{\partial \psi(\theta)}{\partial \theta^i}. \tag{75}$$

The Fisher information matrix is:

$$\begin{aligned} I_{ij}(\theta) &= \mathrm{E}\left[\frac{\partial \log p(X;\theta)}{\partial \theta^i}\frac{\partial \log p(X;\theta)}{\partial \theta^j}\right] && (76)\\ &= \mathrm{E}\left[\left(F_i(X) - \frac{\partial \psi(\theta)}{\partial \theta^i}\right)\left(F_j(X) - \frac{\partial \psi(\theta)}{\partial \theta^j}\right)\right] && (77)\\ &= \mathrm{E}[F_i(X)F_j(X)] - \mathrm{E}[F_i(X)]\frac{\partial \psi(\theta)}{\partial \theta^j} - \mathrm{E}[F_j(X)]\frac{\partial \psi(\theta)}{\partial \theta^i} + \frac{\partial \psi(\theta)}{\partial \theta^i}\frac{\partial \psi(\theta)}{\partial \theta^j}. && (78) \end{aligned}$$

For exponential families, we have:

$$\mathrm{E}[F_i(X)] = \frac{\partial \psi(\theta)}{\partial \theta^i} = \eta_i, \tag{79}$$

where η are the expectation parameters. Therefore:

$$\begin{aligned} I_{ij}(\theta) &= \mathrm{E}[F_i(X)F_j(X)] - \eta_i\eta_j - \eta_j\eta_i + \eta_i\eta_j && (80)\\ &= \mathrm{E}[F_i(X)F_j(X)] - \eta_i\eta_j && (81)\\ &= \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}, && (82) \end{aligned}$$

where the last equality follows from the fact that ψ is the cumulant generating function, so its second derivatives give the covariance. □

A.2 Neural Coding

Proof of Optimal Stimulus Directions. The discriminability between two nearby stimuli s and $s + ds$ is given by the Fisher-Rao distance:

$$d(s, s + ds) = \sqrt{ds^T I(s) ds}. \tag{83}$$

To find the directions that maximize discriminability for a fixed stimulus change magnitude $\|ds\|$, we maximize $ds^T I(s) ds$ subject to $\|ds\|^2 = 1$.

This is a standard quadratic optimization problem. The maximum is achieved when ds is an eigenvector of $I(s)$ corresponding to the largest eigenvalue. The maximum value is $\lambda_{\max}$, the largest eigenvalue.

More generally, if we want to find k orthogonal directions that maximize discriminability, we take the k eigenvectors corresponding to the k largest eigenvalues. These directions are orthogonal with respect to both the Euclidean metric and the Fisher information metric (since $I(s)$ is symmetric and positive definite). □

Proof of Information-Limiting Correlations. For a neural population with mean responses $\boldsymbol{\mu}(s)$ and noise covariance Σ, the Fisher information is:

$$I(s) = \left(\frac{\partial \boldsymbol{\mu}}{\partial s}\right)^T \Sigma^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial s}\right). \tag{84}$$

Let $\mathbf{v} = \partial\boldsymbol{\mu}/\partial s$ be the signal direction. If the principal eigenvector of Σ is parallel to $\mathbf{v}$, then we can write:

$$\Sigma = \sigma^2 \mathbf{v}\mathbf{v}^T + \Sigma_\perp, \tag{85}$$

where $\Sigma_\perp$ has no component in the $\mathbf{v}$ direction.

The Fisher information becomes:

$$I(s) = \mathbf{v}^T \Sigma^{-1} \mathbf{v} = \mathbf{v}^T \left(\frac{1}{\sigma^2}\mathbf{v}\mathbf{v}^T + \Sigma_\perp^{-1}\right) \mathbf{v} = \frac{\|\mathbf{v}\|^2}{\sigma^2}, \tag{86}$$

since $\Sigma_\perp^{-1}\mathbf{v} = 0$ (as $\mathbf{v}$ is orthogonal to the range of $\Sigma_\perp$).

This shows that $I(s) \leq \|\mathbf{v}\|^2/\sigma^2$, regardless of the number of neurons. The bound is achieved when all noise variance is along the signal direction, making the correlations information-limiting. □

A.3 Gradient Learning

Proof of Optimality of Natural Gradient. We want to find the direction $d\theta$ that minimizes the loss function $\mathcal{L}(\theta + d\theta)$ subject to a constraint on the distance traveled on the manifold.

Using the Fisher-Rao metric, the distance constraint is:

$$d\theta^T I(\theta) d\theta = \epsilon^2, \tag{87}$$

for some small $\epsilon > 0$.

The first-order change in the loss is:

$$\mathcal{L}(\theta + d\theta) - \mathcal{L}(\theta) \approx \nabla_\theta \mathcal{L}(\theta)^T d\theta. \tag{88}$$

To minimize this subject to the constraint, we use Lagrange multipliers:

$$L = \nabla_\theta \mathcal{L}(\theta)^T d\theta + \lambda(d\theta^T I(\theta) d\theta - \epsilon^2). \tag{89}$$

Taking the gradient with respect to $d\theta$ and setting it to zero:

$$\nabla_\theta \mathcal{L}(\theta) + 2\lambda I(\theta) d\theta = 0, \tag{90}$$

which gives:

$$d\theta = -\frac{1}{2\lambda} I(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta). \tag{91}$$

This is proportional to the natural gradient direction. The constant factor can be absorbed into the learning rate, proving that the natural gradient is the direction of steepest descent on the manifold. □

Proof of Relationship Between Hessian and Fisher Information. For a loss function $\mathcal{L}(\theta) = -\log p(\mathbf{x}|\theta)$ (negative log-likelihood), the Hessian is:

$$H_{ij}(\theta) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta^i \partial\theta^j} = -\frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial\theta^i \partial\theta^j}. \tag{92}$$

Taking the expectation with respect to $p(\mathbf{x}|\theta)$:

$$\begin{aligned}
\mathrm{E}[H_{ij}(\theta)] &= -\int p(\mathbf{x}|\theta) \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial\theta^i \partial\theta^j} d\mathbf{x} && (93)\\
&= -\int p(\mathbf{x}|\theta) \frac{\partial}{\partial\theta^j} \left(\frac{1}{p(\mathbf{x}|\theta)} \frac{\partial p(\mathbf{x}|\theta)}{\partial\theta^i} \right) d\mathbf{x} && (94)\\
&= -\int \frac{\partial^2 p(\mathbf{x}|\theta)}{\partial\theta^i \partial\theta^j} d\mathbf{x} + \int \frac{1}{p(\mathbf{x}|\theta)} \frac{\partial p(\mathbf{x}|\theta)}{\partial\theta^i} \frac{\partial p(\mathbf{x}|\theta)}{\partial\theta^j} d\mathbf{x}. && (95)
\end{aligned}$$

The first term vanishes because:

$$\int \frac{\partial^2 p(\mathbf{x}|\theta)}{\partial\theta^i \partial\theta^j} d\mathbf{x} = \frac{\partial^2}{\partial\theta^i \partial\theta^j} \int p(\mathbf{x}|\theta) d\mathbf{x} = \frac{\partial^2}{\partial\theta^i \partial\theta^j} 1 = 0. \tag{96}$$

The second term is:

$$\int \frac{\partial \log p(\mathbf{x}|\theta)}{\partial\theta^i} \frac{\partial \log p(\mathbf{x}|\theta)}{\partial\theta^j} p(\mathbf{x}|\theta) d\mathbf{x} = I_{ij}(\theta), \tag{97}$$

which is the Fisher information matrix. Therefore, $\mathrm{E}[H(\theta)] = I(\theta)$. □

Proof of Geodesic Gradient Descent. On a Riemannian manifold $(\mathcal{M}, g)$, the gradient of a function f is defined by:

$$g(\nabla_g f, X) = X(f) = df(X) \tag{98}$$

for all vector fields X.

In local coordinates, if $X = X^i \frac{\partial}{\partial x^i}$, then:

$$X(f) = X^i \frac{\partial f}{\partial x^i} = g_{ij} X^i (\nabla_g f)^j = g(X, \nabla_g f), \tag{99}$$

which gives $(\nabla_g f)^j = g^{ji} \frac{\partial f}{\partial x^i}$.

For gradient descent, we want to move in the direction of steepest descent. On a manifold, this means following a geodesic in the direction of $-\nabla_g f$.

The exponential map $\exp_p(v)$ gives the point reached by following the geodesic starting at p in direction v for unit time. Therefore, the update:

$$\theta_{t+1} = \exp_{\theta_t}(-\eta \nabla_g \mathcal{L}(\theta_t)) \tag{100}$$

moves along the geodesic in the direction of steepest descent, ensuring that the update remains on the manifold.

For small learning rates, this approximates:

$$\theta_{t+1} \approx \theta_t - \eta \nabla_g \mathcal{L}(\theta_t), \tag{101}$$

but the exact exponential map ensures the update stays on the manifold even for larger learning rates. □

A.4 Feature Learning

Sketch of Geometric Evolution During Learning. As a neural network learns, the representational manifold $\mathcal{M}_\theta$ evolves. The Fisher information metric on this manifold measures how sensitive the representation is to changes in the input.

When the network learns task-relevant features, the representation becomes more aligned with task structure. This means:

1. The Fisher information matrix $I(\theta)$ becomes more aligned with task-relevant directions, meaning its principal eigenvectors point in directions that are useful for the task.
2. The curvature of the manifold decreases in task-relevant directions. This can be seen from the fact that the Hessian of the loss function (which relates to curvature) becomes smaller in directions that are well-learned.
3. The volume of the manifold contracts in task-irrelevant directions and expands in task-relevant directions. This follows from the fact that the network learns to ignore irrelevant variations while preserving relevant ones.

These geometric changes can be quantified by monitoring the eigenvalues and eigenvectors of the Fisher information matrix, the principal curvatures of the manifold, and the volume form as training progresses. □

A.5 Spiking Neural Networks

Proof of Fisher Information for Spike Patterns. For a spiking neural network, the spike pattern distribution $p(\mathbf{S}|s)$ gives the probability of observing spike pattern $\mathbf{S}$ given stimulus s.

The Fisher information is:

$$I(s) = \mathbf{E}\left[\left(\frac{\partial \log p(\mathbf{S}|s)}{\partial s}\right)^2 \middle| s\right]. \tag{102}$$

For Poisson-like spike statistics, where spikes are generated independently (or with known correlations), the log-likelihood can be written as:

$$\log p(\mathbf{S}|s) = \sum_i \log p(s_i|s) + \text{correlation terms}, \tag{103}$$

where s_i are individual spike trains.

The Fisher information then becomes:

$$I(s) = \sum_i \mathrm{E}\left[\left(\frac{\partial \log p(s_i|s)}{\partial s}\right)^2\right] + \text{cross terms from correlations.} \tag{104}$$

The first term depends on how sensitive the mean firing rates are to the stimulus. The correlation terms depend on the structure of spike correlations and how they vary with the stimulus.

For Gaussian approximations to spike statistics, this reduces to the standard formula involving the sensitivity of mean responses and the structure of the covariance matrix. □

These proofs provide the mathematical foundations for the key results stated throughout the notes. Some proofs are sketched where full details would require extensive additional background, but the main ideas and techniques are presented.

References

[Ama83] Shun'ichi Amari. A foundation of information geometry. *Electronics and Communications in Japan (Part I: Communications)*, 66(6):1–10, 1983.

[CB24] Jacob T. Crosser and Braden A. W. Brinkman. Applications of information geometry to spiking neural network behavior. *Physical Review E*, 109(2):024302, February 2024. arXiv:2305.07482 [q-bio].

[CLWC25] Chi-Ning Chou, Hang Le, Yichen Wang, and SueYeon Chung. Feature Learning beyond the Lazy-Rich Dichotomy: Insights from Representational Geometry, July 2025. arXiv:2503.18114 [cs].

[DLM$^+$] Xuehao Ding, Dongsoo Lee, Joshua B Melander, George Sivulka, Surya Ganguli, and Stephen A Baccus. Information Geometry of the Retinal Representation Manifold.

[MCC25] Francesca Mignacco, Chi-Ning Chou, and SueYeon Chung. Nonlinear classification of neural manifolds with contextual information, March 2025. arXiv:2405.06851 [q-bio].

[Nie20] Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, September 2020. arXiv:1808.08271 [cs].

[Rao92] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Breakthroughs in Statistics*, pages 235–247, 1992.

[SHW$^+$24] Li Sun, Zhenhao Huang, Qiqi Wan, Hao Peng, and Philip S. Yu. Spiking Graph Neural Network on Riemannian Manifolds, October 2024. arXiv:2410.17941 [cs] version: 1.

[SR21] Rava Azeredo da Silveira and Fred Rieke. The Geometry of Information Coding in Correlated Neural Populations. *Annual Review of Neuroscience*, 44(1):403–424, July 2021. arXiv:2102.00772 [q-bio].