Mini-Course on Information Geometry

Introduction

Herlock Rahimi
Department of Electrical and Computer Engineering
Yale University

June 5, 2025

Overview

Information

What is Information?

- Information measures how much uncertainty is reduced when we observe an event.
- **Self-information** of an event x with probability $p(x)$:

$$I(x) = -\log p(x).$$

- Interpretation:
  - If $p(x)$ is high, $I(x)$ is low $\Rightarrow$ the event is expected, carries little information.
  - If $p(x)$ is low, $I(x)$ is high $\Rightarrow$ the event is surprising, carries more information.

Entropy: The Expected Information

- Entropy measures the average uncertainty (or surprise) in a probability distribution.
- **Shannon Entropy** for a discrete random variable:

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

- **Differential Entropy** for a continuous random variable:

$$H(X) = -\int p(x) \log p(x) dx.$$

- Entropy is maximized when all outcomes are equally probable.
- Entropy gives a fundamental limit on data compression and transmission.

Why is This Related to Information?

- Entropy quantifies the fundamental limit of how much information is needed to describe a random variable.
- If entropy is high:
  - The distribution is more uniform.
  - More bits are required to describe outcomes.
- If entropy is low:
  - The distribution is concentrated on a few outcomes.
  - Fewer bits are required to describe outcomes.
- Many information measures (KL divergence, mutual information) build on entropy.

Likelihood

Likelihood Function:

- Given data $X = \{x_1, \ldots, x_N\}$ and a parametric model $p(x|\theta)$, the likelihood function is:

$$L(\theta) = \prod_{i=1}^{N} p(x_i|\theta).$$

- Taking the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta).$$

KL Divergence: Measuring Information Difference

- KL divergence measures the difference between two probability distributions $p(x)$ and $q(x)$.
- Definition:

$$D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

- Interpretation:
  - Measures how much information is lost when using $q(x)$ instead of $p(x)$.
  - KL divergence is always non-negative and zero if $p(x) = q(x)$.
  - It is not symmetric: $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$.

Likelihood and Information (Entropy)

Connection to Entropy:

- The **expected log-likelihood** under the true distribution $p(x)$ is:

$$\mathbb{E}_{p(x)}[\log p(x|\theta)] = \int p(x) \log p(x|\theta) dx.$$

- Using entropy $H(p) = -\int p(x) \log p(x) dx$, we can rewrite:

$$\mathbb{E}_{p(x)}[\log p(x|\theta)] = -H(p) - D_{\mathrm{KL}}(p \| p_\theta).$$

- **Maximizing likelihood** is equivalent to minimizing KL divergence between the true and model distributions.

What is Wasserstein Distance?

- Wasserstein distance, also called the Earth Mover's Distance (EMD), measures the effort required to transform one probability distribution into another.
- It is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int |x - y|^p d\gamma(x, y) \right)^{\frac{1}{p}},$$

where:

  - μ, ν are two probability distributions.
  - $\Gamma(\mu, \nu)$ is the set of joint distributions with marginals μ and ν.
  - It represents the optimal transport cost to morph μ into ν.
  - Applications in Finance, Portfolio Optimization, etc.

An Alternative: **Intuition Behind Wasserstein Distance**

- Imagine moving piles of dirt from one location to another.
- Wasserstein distance measures:
  - How much dirt needs to be moved.
  - How far each unit of dirt must travel.
- If two distributions are similar, the transportation cost is small.
- If they are far apart, the cost is large.

Estimation

The Estimation Problem

- Given a dataset $X = \{x_1, x_2, \ldots, x_N\}$, assume observations are drawn from a normal distribution:
$$x_i \sim \mathcal{N}(\mu, \sigma^2).$$
- Goal: Estimate (μ, σ^2) from the observed data.
- We consider three estimation approaches:
  1. **Maximum Likelihood Estimation (MLE)**
  2. **Minimization of KL Divergence with Empirical Distribution**
  3. **Minimization of Wasserstein Distance**

MLE: Maximum Likelihood Estimation

Likelihood function: Given N i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$,

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Log-likelihood:

$$\ell(\mu, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

MLE estimates: Solve

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \sigma^2} = 0.$$

This gives:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2.$$

Minimizing KL Divergence with the Empirical Distribution

- The empirical distribution is defined as:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i),$$

  where $\delta(x - x_i)$ is the Dirac delta function.
- Substituting the empirical distribution:

$$D_{\mathrm{KL}}(\hat{p}\|q) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\delta(x_i - x)}{q(x|\mu, \sigma^2)}.$$

- Minimizing this divergence leads to the same estimates:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2.$$

Minimizing Wasserstein Distance

- Wasserstein-2 distance between two normal distributions:

$$W_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

- Minimizing $W_2^2(\hat{p}, q)$ leads to:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{\mu}|.$$

- This differs from MLE: instead of variance, the Wasserstein estimator uses the mean absolute deviation.

Comparison of KL and Wasserstein Minimization

- Given a dataset $X = \{x_1, \dots, x_N\}$ of size $N = 100$, we estimate a normal distribution $\mathcal{N}(\mu, \sigma^2)$ using:
  1. **Maximum Likelihood Estimation (MLE)** by minimizing KL divergence.
  2. **Optimal Transport Estimation** by minimizing Wasserstein distance.
- The true distribution used to generate the data:

$$X_i \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}}) = \mathcal{N}(2, 1.5^2).$$

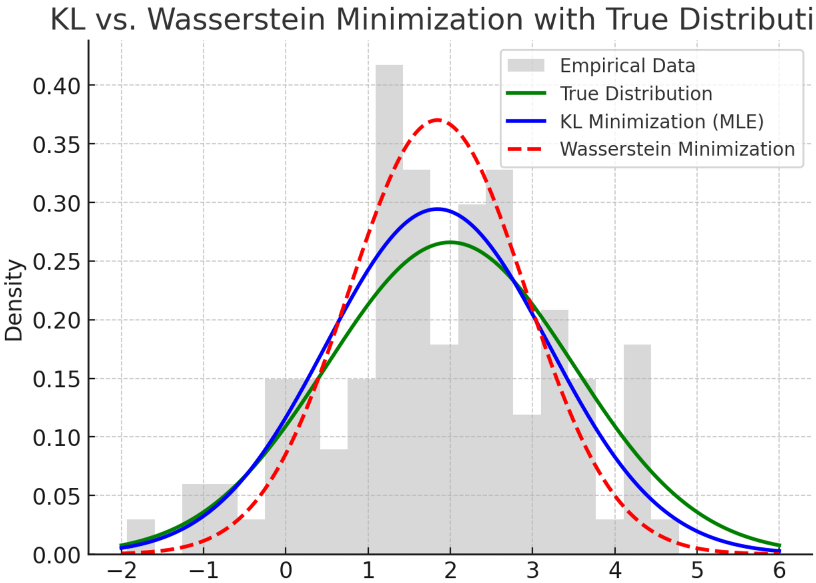
- **MLE (KL Minimization)** results in:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2.$$

- **Wasserstein Distance Minimization** results in:

$$\hat{\mu}_{W_2} = \hat{\mu}_{\text{MLE}}, \quad \hat{\sigma}_{W_2} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{\mu}|.$$

Comparison of KL and Wasserstein Minimization



Conclusion

- Both KL divergence and Wasserstein distance define a notion of "distance" between distributions.
- Minimization of these distances leads to meaningful estimators.
- KL divergence minimize "information" loss.
- Wasserstein distance provides an alternative by minimizing the "movement" needed.

Geometry

Optimization and Geometry

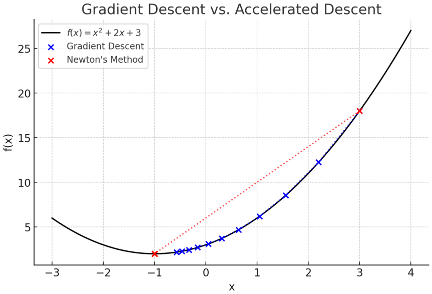
- Optimization problems often involve navigating high-dimensional spaces.
- Standard gradient descent assumes a Euclidean structure.
- But real problems often have underlying curvature!
- Using geometric information (curvature) can significantly speed up convergence.

Example: Gradient Descent on a Quadratic Function

Consider optimizing the function:

$$f(x) = x^2 + 2x + 3$$

- Standard gradient descent takes slow steps along the gradient.
- Second-order information (curvature) helps adjust steps, leading to accelerated descent.

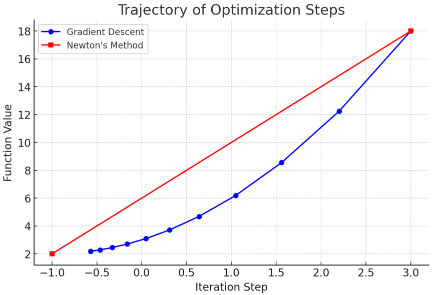


Geometric Acceleration: Hessian and Curvature

- The second derivative (Hessian) captures curvature.
- Newton's method uses this information:

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

- Leads to faster convergence as seen in the trajectory plot.



Conclusion

- Geometry plays a key role in optimization.
- Using curvature information can drastically improve convergence.
- Information Geometry extends these ideas to probability distributions and statistical manifolds.

Example: Optimization on a Torus

Optimization Problem:

- We aim to minimize a function $f(x, y, z)$ constrained to a toroidal surface.
- Let the torus be defined by:

$$(R - \sqrt{x^2 + y^2})^2 + z^2 = r^2,$$

  where R is the major radius and r is the minor radius.
- Consider a simple quadratic objective function:

$$f(x, y, z) = (x - x^*)^2 + (y - y^*)^2 + (z - z^*)^2,$$

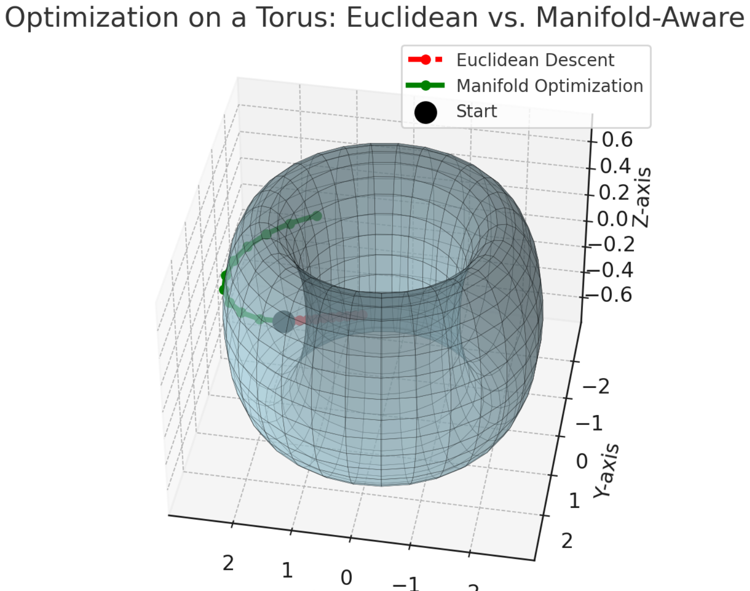
  where (x^*, y^*, z^*) is a target point on the torus.

Example: Optimization on a Torus

Comparison of Optimization Approaches:

- **Euclidean Gradient Descent:**
  - Ignores the toroidal constraint.
  - Moves in the naive gradient direction.
  - Results in inefficient steps off the toroidal structure.
- **Manifold-Aware Optimization:**
  - Moves along the geodesics of the torus.
  - Preserves feasibility within the constraint.
  - Converges faster with more efficient steps.

Example: Optimization on a Torus



Optimization and Geometry: The Fundamental Connection

- Many optimization problems involve navigating a high-dimensional space.
- The structure of this space is often **curved**, rather than flat.
- Standard gradient descent assumes a **Euclidean structure**, but real-world problems often have curvature.
- Taking geometry into account can significantly improve optimization speed and accuracy.

Riemannian Manifold

What is a Differentiable Manifold?

- A differentiable manifold M of dimension n is a topological space that is locally homeomorphic to $\mathbb{R}^n$ and has a smooth structure.
- Formally, a differentiable manifold consists of:
  - A set M.
  - A collection of charts $(U_\alpha, \varphi_\alpha)$, where $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$ is a homeomorphism.
  - Transition functions $\varphi_\beta \circ \varphi_\alpha^{-1}$ that are smooth wherever defined.
- Example: The n-dimensional sphere S^n is a differentiable manifold.

Tangent Vectors and Tangent Space

- A tangent vector at a point $p \in M$ is an equivalence class of smooth curves $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$.
- The tangent space at p, denoted T_pM, is the space of all tangent vectors at p.
- If $(x^1, \ldots, x^n)$ are local coordinates, then a tangent vector is expressed as:

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

- Example: The tangent space at a point on the 2-sphere S^2 consists of all vectors tangent to the sphere at that point.

Tensor Fields

- A tensor of type (r, s) at a point $p \in M$ is a multilinear map:

$$T : (T_p M^*)^r \times (T_p M)^s \to \mathbb{R}.$$

- A tensor field assigns a tensor T_p to each point p in a smooth manner.
- Example: The metric tensor g is a tensor field of type $(0, 2)$ that defines an inner product on each $T_p M$.
- In local coordinates $(x^1, \ldots, x^n)$, a tensor field is written as:

$$T = T^{i_1 \ldots i_r}_{j_1 \ldots j_s} \frac{\partial}{\partial x^{i_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{i_r}} \otimes dx^{j_1} \otimes \cdots \otimes dx^{j_s}.$$

Riemannian Metrics and Inner Products

- A Riemannian metric on a manifold M is a smoothly varying positive-definite inner product g_p on each tangent space T_pM.
- That is, for every $p \in M$ and $v, w \in T_pM$, we have:

$$g_p(v, w) = g_{ij}(p) v^i w^j.$$

- The Riemannian metric allows us to define:
  - Length of vectors: $\|v\| = \sqrt{g_p(v, v)}$.
  - Angle between vectors.
  - Distance between points on M as:

$$d(p, q) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt.$$

- Example: The standard metric on $\mathbb{R}^n$ is the Euclidean metric $g_{ij} = \delta_{ij}$, while the metric on the sphere is the induced metric from $\mathbb{R}^{n+1}$.

Summary of Differential Geometry Essentials

- Differentiable Manifolds: Generalize smooth surfaces to higher dimensions.
- Tangent Vectors and Tangent Space: Capture local directions at each point.
- Tensor Fields: Generalize scalars, vectors, and covectors.
- Riemannian Metrics: Define distances, angles, and inner products on manifolds.

Distance on a Riemannian Manifold

- A Riemannian manifold (M, g) is a smooth manifold M equipped with a Riemannian metric g.
- The Riemannian distance between two points $p, q \in M$ is defined as:

$$d(p, q) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt,$$

  where the infimum is taken over all smooth curves γ connecting p and q.
- The choice of metric tensor g determines the geometry of the space.
- We consider two fundamental geometries:
  - Fisher-Rao Geometry (Induced by KL Divergence)
  - Wasserstein Geometry (Induced by Optimal Transport)

Information Geometry

Information Geometry

- Geometry is essential for Optimization.
- To induce Geometry we need a metric. The metric shows what is important for us.

The One-Dimensional Normal Distribution as a Riemannian Manifold

Manifold Structure:

- The family of one-dimensional normal distributions:

$$\mathcal{N}(\mu, \sigma^2) = \left\{ p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right\}$$

  forms a 2-dimensional Riemannian manifold with coordinates (μ, σ).
- The natural Riemannian structure depends on the choice of the metric tensor g_{ij}.

Distance on a Riemannian Manifold:

- A Riemannian metric g defines infinitesimal distances:

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

- We consider two choices of metric:
  1. **Fisher Information Metric** (Statistical Information Geometry)
  2. **Optimal Transport Metric** (Wasserstein Geometry)

Fisher Information Metric for Normal Distributions

Fisher Information Matrix:

- The Fisher information metric is derived from the Fisher information matrix:

$$g_{ij}(\mu, \sigma) = \mathbb{E}\left[\frac{\partial \log p(x|\mu, \sigma)}{\partial \theta^i} \frac{\partial \log p(x|\mu, \sigma)}{\partial \theta^j}\right].$$

- Computing the derivatives:

$$\frac{\partial \log p(x|\mu, \sigma)}{\partial \mu} = \frac{x - \mu}{\sigma^2}, \quad \frac{\partial \log p(x|\mu, \sigma)}{\partial \sigma} = \frac{(x - \mu)^2 - \sigma^2}{\sigma^3}.$$

- Taking expectations over $p(x|\mu, \sigma)$, the Fisher information matrix is:

$$g_{\text{Fisher}}(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}.$$

Wasserstein Metric for Normal Distributions

Optimal Transport (Wasserstein-2) Distance:

- The Wasserstein-2 distance between two normal distributions:

$$W_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

- This induces a Riemannian metric:

$$g_{\text{Wass}}(\mu, \sigma) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Comparison with Fisher Metric:

- The Wasserstein metric treats (μ, σ) as Euclidean parameters, leading to a flat geometry.
- The Fisher metric accounts for statistical structure, leading to non-Euclidean curvature.
- The geodesics in Fisher-Rao space differ from those in Wasserstein space.

Estimating Normal Distribution Parameters

Goal: Estimate parameters (μ, σ) of a normal distribution from a dataset.

- Assume data $X = \{x_1, x_2, ..., x_{100}\}$ comes from:

$$X_i \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}}).$$

- Given X, we estimate (μ, σ) via gradient descent:
  1. **Wasserstein Gradient Descent** (which reduces to Euclidean).
  2. **Natural Gradient Descent** using Fisher Information.
  3. We start the estimation from $(\mu, \sigma) = (1, 3)$

Example Setup:

- $\mu_{\text{true}} = 3$, $\sigma_{\text{true}} = 1$.
- 100 samples from $\mathcal{N}(2, 1.5^2)$.

Wasserstein Gradient Descent (Euclidean)

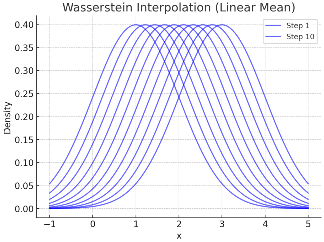
- The Wasserstein-2 metric for normal distributions:

$$W_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

- This reduces to Euclidean gradient descent in parameter space:

$$\mu_{t+1} = \mu_t - \eta \frac{\partial L}{\partial \mu}, \quad \sigma_{t+1} = \sigma_t - \eta \frac{\partial L}{\partial \sigma}.$$

- This method treats the parameter space as a flat Euclidean space.



Natural Gradient Descent (Fisher Information)

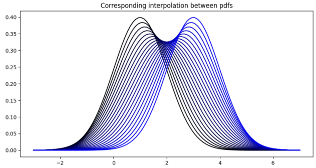
- The Fisher Information Metric for a normal distribution is:

$$g(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}.$$

- The natural gradient descent update is:

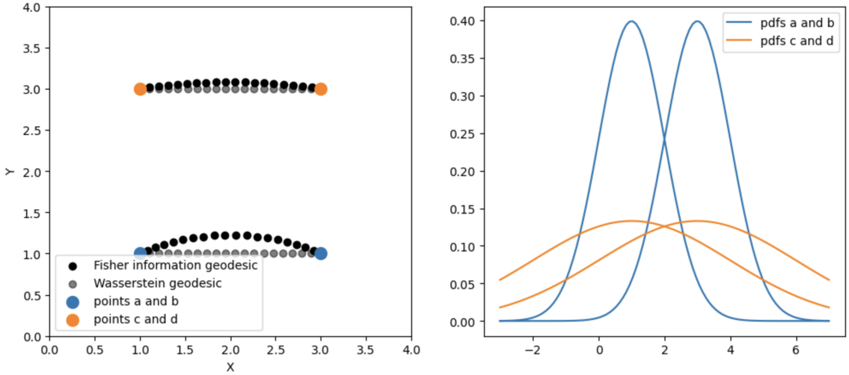
$$\theta_{t+1} = \theta_t - \eta g(\theta)^{-1} \nabla_\theta L.$$

- Unlike Euclidean descent, this accounts for the curvature of the parameter space.



Comparison: Wasserstein vs. Natural Gradient Descent

- Wasserstein gradient descent (Euclidean) updates parameters slowly.
- Fisher natural gradient descent accounts for curvature and converges faster.
- Key Result: The Fisher-Rao metric rescales gradients, making steps more efficient.



Conclusion: Why Fisher Natural Gradient is Faster?

- Wasserstein descent treats the parameter space as Euclidean, leading to slow updates.
- Fisher natural gradient descent corrects for parameter space curvature.
- **Natural gradient updates are more efficient and converge faster.**

Conclusion

- What is Information?
- What is Geometry?
- Why Geometric Approach toward Information is necessary(**Information Geometry**).

Next time:

- we would discuss in more details the mathematics of Differential Geometry.
- How one can tackle Problems in Machine Learning and Finance with Geometric Approaches.

References

- Amari, S. (2000). *Methods of Information Geometry*. American Mathematical Society.
- Amari, S. (2016). *Information Geometry and Its Applications*. Springer.
- Lee, J. M. (2018). *Introduction to Riemannian Manifolds*. Springer.
- do Carmo, M. (1992). *Riemannian Geometry*. Birkhäuser.
- Gallot, S., Hulin, D., & Lafontaine, J. (2004). *Riemannian Geometry*. Springer.
- Pennec, X. (2020). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press.

Question?