Mini-Course on Information Geometry

Introduction

Herlock Rahimi
Department of Electrical and Computer Engineering
Yale University

June 5, 2025

Overview

1. Optimal Transport
2. Information Geometry meets Optimal Transport
3. Information Geometry of Risk And Returns

What is Optimal Transport?

Objective: Find the most efficient way to transport mass from one distribution to another.

- Given: two probability measures μ and ν defined on measurable spaces X and Y, respectively.
- Goal: Move the entire mass of μ to match the distribution ν with minimal cost.
- Cost: A function $c : X \times Y \to \mathbb{R}$ encoding the cost of moving a unit mass from $x \in X$ to $y \in Y$.

We will examine two formulations: Monge and Kantorovich.

Monge Formulation (1781)

Transport via deterministic map:

- $T : X \to Y$ is a measurable map.
- T pushes μ forward to ν: $T_{\#}\mu = \nu$ means $\nu(B) = \mu(T^{-1}(B))$ for all Borel sets $B \subseteq Y$.

Monge's Problem: Find T minimizing total transport cost:

$$\inf_{T : T_{\#}\mu = \nu} \int_X c(x, T(x))\, d\mu(x)$$

Issues:

- Not all ν can be written as pushforwards of μ.
- No mass splitting: each x maps to one y.
- Highly nonlinear; existence is not guaranteed.

Kantorovich Formulation (1940s)

Key idea: Allow mass to split by using transport plans.

- A **coupling** or **transport plan** π is a probability measure on $X \times Y$.
- π must have marginals μ and ν:

$$\int_Y d\pi(x, y) = \mu(x) \quad \text{(first marginal)}$$

$$\int_X d\pi(x, y) = \nu(y) \quad \text{(second marginal)}$$

- The space of admissible plans is denoted $\Pi(\mu, \nu)$.

Minimize:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

This is a linear program over the space of couplings.

Duality in Kantorovich Problem

Convex dual formulation:

$$\sup_{\varphi \in C(X), \psi \in C(Y)} \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \;\middle|\; \varphi(x) + \psi(y) \leq c(x,y) \right\}$$

Key functions:

- $\varphi : X \to \mathbb{R}$, $\psi : Y \to \mathbb{R}$
- They are dual to each other under the c-**transform**:

$$\varphi^c(y) = \inf_{x \in X} [c(x,y) - \varphi(x)]$$

- A function is c-**convex** if $\varphi = (\varphi^c)^c$

Analogy: c-convexity generalizes Legendre-Fenchel duality.

Existence and Structure of Optimal Maps

Assumptions:

- $X = Y = \mathbb{R}^n$, $c(x, y) = \|x - y\|^2$, and μ is absolutely continuous.

Then: There exists a convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that the optimal map is:

$$T(x) = \nabla\varphi(x)$$

More generally:

- $T(x) = c\text{-}\exp_x(\nabla\varphi(x))$ where:

$$c\text{-}\exp_x(p) = y \iff p = -\nabla_x c(x, y)$$

Domain: $\nabla\varphi : \mathbb{R}^n \to \mathbb{R}^n$

Jacobian condition: for optimal map $T = \nabla\varphi$:

$$\det(DT(x)) = \frac{d\mu(x)}{d\nu(T(x))}$$

In PDE form:

$$\det D^2\varphi(x) = \frac{d\mu(x)}{d\nu(\nabla\varphi(x))}$$

This is the Monge-Ampère equation, a highly nonlinear elliptic PDE.

- $D^2\varphi$ is the Hessian (matrix of second derivatives)

Wasserstein Distance: A Metric on Distributions

Setup: Let μ, ν be two probability measures on $\mathbb{R}^n$ with finite p-th moments.

- Let $\Pi(\mu, \nu)$ denote the set of couplings π on $\mathbb{R}^n \times \mathbb{R}^n$ such that:

$$\pi(A \times \mathbb{R}^n) = \mu(A)$$
$$\pi(\mathbb{R}^n \times B) = \nu(B)$$

- Think of $\pi(x, y)$ as describing how much mass is transported from x to y.

Definition: Wasserstein-p distance between μ and ν:

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^p \, d\pi(x, y) \right)^{1/p}$$

Properties:

- W_p defines a metric on the space $\mathcal{P}_p(\mathbb{R}^n)$ of probability measures with finite p-th moment.
- W_2 is the most widely used in analysis and geometry due to its deep structure.
- Encodes geometric information about how "far apart" distributions are.

Displacement Interpolation: Geodesics in $\mathcal{P}_2$

Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^n)$ and $T = \nabla\varphi$ be the optimal map sending μ_0 to μ_1.

- We define a path of measures interpolating between μ_0 and μ_1:

$$\mu_t := ((1-t)\mathrm{Id} + tT)_{\#}\mu_0, \quad t \in [0,1]$$

- This means each point x in the support of μ_0 is transported to $T(x)$ along a straight line, at time t it is at $(1-t)x + tT(x)$.

Interpretation:

- The curve $t \mapsto \mu_t$ is a **geodesic** in Wasserstein space $\mathcal{P}_2$.
- This is not a pointwise interpolation, but a mass-preserving geodesic.
- Interpolated densities may be nontrivial even when endpoints are singular.

Otto Calculus: A Riemannian View of $\mathcal{P}_2$

Felix Otto (2001): Interpreted $\mathcal{P}_2(\mathbb{R}^n)$ as an infinite-dimensional Riemannian manifold.
Tangent space: For μ with smooth positive density ρ, the tangent space is:

$$T_\mu \mathcal{P}_2 = \{v : \partial_t \rho_t + \nabla \cdot (\rho_t v) = 0\}$$

- This is derived from the continuity equation — conservation of mass.

Riemannian metric: The inner product on $T_\mu \mathcal{P}_2$ is:

$$\langle v, w \rangle_{T_\mu} := \int_{\mathbb{R}^n} \langle v(x), w(x) \rangle \, d\mu(x)$$

This structure allows defining gradients and geodesics on $\mathcal{P}_2$ as in finite-dimensional manifolds.

Gradient Flows in $\mathcal{P}_2$: The Fokker–Planck Equation

Consider functional: $\mathcal{F}(\mu) = \int \rho(x) \log \rho(x)\, dx$ (entropy)

- Gradient flow of $\mathcal{F}$ in $\mathcal{P}_2$ leads to:

$$\partial_t \rho = \nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$$

- This is the heat equation — a fundamental diffusion process.

Otto's insight: Many nonlinear PDEs can be seen as gradient flows in Wasserstein geometry.

- This viewpoint unifies diffusion, fluid dynamics, and thermodynamics.

Extended Summary of Section 1

- **Wasserstein metric:** Quantifies distance between distributions using cost-minimizing transport.
- **Displacement interpolation:** Describes geodesic paths in the space of probability measures.
- **Otto calculus:** Provides differential geometry tools for probability spaces.
- **Gradient flows:** Classical PDEs arise as natural flows in the metric geometry of $\mathcal{P}_2$.

Conclusion: Optimal transport provides a geometric and analytical foundation for studying distributional evolution. **Next:** Section 2 — introducing information geometry and its synergy with optimal transport.

We compare:

- **Wasserstein geometry:** Arises from optimal transport theory.
- **Information geometry:** Arises from statistics and information theory.

Both geometries live on the space of probability distributions, but are fundamentally different:

- They induce different notions of distance, geodesics, curvature, and gradients.
- They arise from different optimization principles and have different applications.

Wasserstein Geometry Recap

Foundation: Cost-minimizing transport between probability distributions.

- Metric: $W_2(\mu, \nu)$ measures minimum transport effort.
- Geodesics: Displacement interpolations μ_t defined by optimal maps.
- Gradient flows: Derived from mass-preserving PDEs (e.g., heat equation).
- Curvature: Can encode displacement convexity, Ricci curvature bounds (Lott-Sturm-Villani).

Geometry: Riemannian, infinite-dimensional, grounded in mass conservation.

Information Geometry Recap

Foundation: Parametric statistical models and divergence measures.

- Metric: Fisher information metric $g_{ij}(\theta) = \mathbb{E}\left[\partial_i \log p(x;\theta)\partial_j \log p(x;\theta)\right]$
- Distance: No true metric, but divergences (e.g., KL divergence) act as pseudo-distances.
- Geodesics: Exponential (e-) and mixture (m-) families with dual connections.
- Gradient flows: Natural gradient descent, mirror descent.

Geometry: Dually flat affine geometry, finite-dimensional, built from divergences.

Geometric Structures Compared

Metric Structures:

- Wasserstein: W_2 is a true distance; has associated Riemannian structure.
- Information geometry: Fisher metric is Riemannian, but distances often via divergences (e.g., KL).

Geodesics:

- Wasserstein: Displacement geodesics move mass.
- Information geometry: Exponential and mixture geodesics change parameters.

Curvature:

- Wasserstein: Variable curvature (e.g. Ricci bounds in LSV theory).
- Information geometry: Flat under dual affine connections.

Underlying Principles

Optimization Foundations:

- Wasserstein: Cost minimization in transporting mass.
- Information geometry: Divergence minimization (e.g., KL) for estimation/inference.

Infinitesimal Distances:

- Wasserstein: Quadratic cost of displacing mass: $\delta^2 = \int \|v\|^2 \, d\mu$.
- Information geometry: Infinitesimal change in KL divergence: $\delta^2 = \int (\nabla \log p)^2 \, dp$.

Gradient Flow Examples:

- Wasserstein: Entropy gradient flow gives heat equation.
- Information geometry: Gradient flow of KL gives natural gradient descent.

Summary: Complementary Geometries

Wasserstein geometry:

- Focused on *moving mass*, grounded in transport.
- Emphasizes physical processes, PDEs, fluid flows.

Information geometry:

- Focused on *changing beliefs*, grounded in inference.
- Emphasizes statistical structure, duality, and learning.

Goal of the paper: Combine these to form a unified geometry of distributional dynamics.

Core question: How can the Monge problem in optimal transport be connected to the Fisher information geometry?

- **Monge OT:** Find the optimal transport map T pushing μ to ν minimizing cost.
- **Fisher geometry:** Defined on statistical manifolds with Riemannian structure from Fisher information.
- This section draws a bridge: views OT as a Riemannian geometry with deep links to information theory.

Setup: Probability Measures as Points on a Manifold

Let $\mathcal{P}_+(\Omega)$ denote the space of smooth, strictly positive probability densities on domain $\Omega \subseteq \mathbb{R}^n$.

- Each $\rho \in \mathcal{P}_+(\Omega)$ satisfies: $\rho(x) > 0$ and $\int_\Omega \rho(x)dx = 1$.
- The tangent space $T_\rho \mathcal{P}_+$ consists of functions σ satisfying $\int \sigma = 0$ (preserves total mass).

Observation: We can endow this manifold with either:

- the **Fisher metric** from statistics, or
- the **Wasserstein metric** from transport.

Fisher Metric as a Riemannian Metric

Fisher metric at ρ:

$$g_\rho^F(\sigma_1, \sigma_2) = \int \frac{\sigma_1(x)\sigma_2(x)}{\rho(x)} dx$$

Interpretation:

- Measures squared fluctuations of the score function.
- Arises from the second-order expansion of KL divergence.
- Used in natural gradient descent, information projections, statistical inference.

It defines a flat Riemannian structure on statistical manifolds.

Wasserstein Metric via Benamou–Brenier

Wasserstein metric at ρ:

$$g_\rho^W(\sigma_1, \sigma_2) = \int \nabla\phi_1(x) \cdot \nabla\phi_2(x)\rho(x)dx$$

where $\sigma_i = -\nabla \cdot (\rho\nabla\phi_i)$

- Encodes transport cost by solving Poisson equations for potentials.
- Geometry depends on gradient flows — velocity fields move mass.
- Fisher metric uses functions on density space; Wasserstein metric lifts to vector fields.

Why Compare Monge and Fisher?

Both geometries are Riemannian — but on different terms:

- Fisher: Intrinsic geometry from divergence minimization (infinitesimal KL).
- Monge: Extrinsic geometry from minimal displacement cost (quadratic effort).

Insight: The OT view (Monge/Wasserstein) captures
dynamic transport behavior — Fisher geometry does not.

- But Fisher has deep ties to statistical structure.
- Goal: bridge the two by finding transport analogues of Fisher structures.

Summary: Geometrizing Probability

- The space of probability densities can carry different Riemannian structures:
  - Fisher: from statistical inference and information theory.
  - Wasserstein: from optimal transport and mechanics.
- The Monge formulation leads to a dynamical geometric view — the Fisher metric is static.
- The next sections will examine how these geometries interact and can be unified.

Next: Comparing their flows, divergences, and potential integration.

Goal: Compare divergence measures that induce geometry on spaces of probability distributions.

- **In Information Geometry:** Divergences generate metrics and dual connections.
- **In Optimal Transport:** Divergences arise as dynamic costs or convex functionals.

We focus on:

- The Kullback–Leibler (KL) divergence
- The Wasserstein distance (as a functional)
- The entropy and its role as a generating functional

KL Divergence and Fisher Metric

Definition (Relative Entropy):

$$D_{\mathrm{KL}}(\rho \| \nu) = \int \rho(x) \log \left(\frac{\rho(x)}{\nu(x)} \right) dx$$

Properties:

- Non-negative, convex, equals 0 iff $\rho = \nu$.
- **Second-order expansion:** around ν gives Fisher metric.

Geometric role: KL acts as a local squared distance:

$$D_{\mathrm{KL}}(\rho + \varepsilon\sigma \| \rho) = \frac{\varepsilon^2}{2} \int \left(\frac{\sigma(x)}{\rho(x)} \right)^2 \rho(x) dx + o(\varepsilon^2)$$

This quadratic form defines the Fisher information.

Wasserstein Distance as Functional

Wasserstein-2 squared distance:

$$W_2^2(\rho, \nu) = \inf_{T : T_{\#}\rho = \nu} \int \|x - T(x)\|^2 \, d\rho(x)$$

Interpreted as: minimal kinetic energy of moving mass from ρ to ν.

- Encodes dynamic effort, not divergence in density values.
- Can be viewed as a squared distance functional on the manifold $\mathcal{P}_2$.

In Otto calculus: W_2^2 plays the role of squared Riemannian distance.

Entropy as a Generating Functional

Boltzmann–Shannon Entropy:

$$\mathcal{H}(\rho) = \int \rho(x) \log \rho(x)\, dx$$

In Information Geometry:

- Negative entropy generates the KL divergence.
- $\nabla \mathcal{H}(\rho) = 1 + \log \rho(x)$ is the natural parameter.

In Optimal Transport:

- Entropy gradient flow in W_2 geometry gives the heat equation.
- Plays the role of a convex potential in variational transport problems.

Conclusion: Entropy unifies divergence-based and transport-based variational principles.

Three Divergences Compared

KL Divergence:

- Asymmetric; measures informational discrepancy.
- Generates Fisher metric via local expansion.

Wasserstein Distance:

- Symmetric (in W_2); measures physical cost of rearrangement.
- Generates transport-based geometry.

Entropy Functional:

- Appears in both settings: as divergence generator and flow potential.
- Central to variational formulations.

Summary: Divergence Functionals and Geometry

- Divergences define geometric structures: metric, connections, and flows.
- KL divergence yields the Fisher metric and underlies statistical estimation.
- Wasserstein distance defines geometry from dynamics and mechanics.
- Entropy is the key potential bridging both settings.

Next: Unifying the geometry via flows and second-order structures.

Section 2.4: Gradient Flows in Probability Space

Goal: Understand how different geometries induce different gradient flows.

- Gradient flows describe the evolution of distributions to minimize a functional.
- The choice of geometry (Wasserstein or Fisher) determines the form of this flow.
- We compare flows derived from entropy in both settings.

Let $\mathcal{F}(\rho)$ be a functional on densities (e.g., KL divergence).

- Fisher geometry defines steepest descent as:

$$\partial_t \rho = -\mathrm{grad}^F \mathcal{F}(\rho)$$

- In coordinates:

$$\partial_t \rho = -\nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right)$$

For entropy: $\mathcal{F}(\rho) = \int \rho \log \rho \Rightarrow$ heat equation:

$$\partial_t \rho = \Delta \rho$$

Flow is conservative and information-theoretic.

Gradient Flow under Wasserstein Geometry

Otto calculus defines gradient flow in W_2 as:

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right)$$

- This is the opposite sign of Fisher flow — consistent with steepest descent in Wasserstein space.
- It encodes the velocity field derived from minimizing transport effort.

Entropy flow:

$$\mathcal{F}(\rho) = \int \rho \log \rho \Rightarrow \partial_t \rho = \Delta \rho$$

Same PDE — different geometry and interpretation.

Interpretation of Gradient Flow Duality

Why do both geometries lead to the heat equation from entropy?

- Because entropy is **convex in both geometries**.
- Gradient flow = steepest descent of a convex functional.
- The **velocity field vs. functional derivative** perspective separates Wasserstein and Fisher.

Key distinction:

- Fisher: views $\nabla \log \rho$ as a statistical object (score function).
- Wasserstein: interprets $\nabla \log \rho$ as a transport velocity field.

Summary: Geometry Determines Flow

- Gradient flow = steepest descent in chosen geometry.
- Fisher gradient: leads to flows via divergence minimization.
- Wasserstein gradient: leads to flows via dynamic transport.
- Entropy yields heat equation under both, revealing deep compatibility.

Next: Second-order geometry — how curvature and acceleration emerge from these flows.

Section 2.6: Entropy-Regularized Optimal Transport

Motivation: Classical OT is computationally expensive — especially in high dimensions.

- The OT problem is a linear program: costly and unstable numerically.
- Adding entropy regularization smooths the problem.
- Leads to faster and more robust algorithms.

Intuition: The regularized problem prefers couplings with higher entropy — spreads mass more evenly.

The Entropy-Regularized Problem

Kantorovich formulation with entropy penalty:

$$\pi^{\varepsilon} = \arg \min_{\pi \in \Pi(\mu,\nu)} \int c(x,y)d\pi(x,y) + \varepsilon D_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

- $D_{\mathrm{KL}}(\pi \| \mu \otimes \nu) = \int \log \left(\frac{d\pi}{d\mu \otimes d\nu} \right) d\pi$
- $\varepsilon > 0$ controls the strength of smoothing.
- As $\varepsilon \to 0$, the solution converges to true OT.

Solution: Gibbs Kernel and Sinkhorn Algorithm

Optimal solution: Takes the form:

$$\pi^{\varepsilon}(x,y) = u(x)K(x,y)v(y), \quad K(x,y) = e^{-c(x,y)/\varepsilon}$$

- u, v are scaling functions determined iteratively.
- Algorithm: Sinkhorn iterations alternate between normalizing rows and columns.

Computational benefits:

- Turns OT into matrix scaling.
- Logarithmic convergence; GPU-efficient.

Geometric Interpretation

Entropy regularization changes the geometry:

- Adds a strongly convex term to the OT objective.
- Implies a smoothed version of the Wasserstein distance.
- Can be viewed as interpolation between OT and KL geometry.

This smoothness improves numerical stability and differentiability.

Summary: Why Entropic OT is Important

- **Practical:** Computable via Sinkhorn scaling — fast and scalable.
- **Theoretical:** Interpolates between geometry of OT and information divergence.
- **Conceptual:** Regularization unifies transport and statistical entropy.

Conclusion: Entropic OT is central in modern computational OT and variational inference.

Sinkhorn Algorithm for Entropic OT

Problem: Compute the entropy-regularized optimal transport plan

$$\pi^\varepsilon = \arg\min_{\pi \in \Pi(\mu,\nu)} \int c(x,y)d\pi(x,y) + \varepsilon D_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

Key object: Gibbs kernel

$$K_{ij} = \exp\left(-\frac{C_{ij}}{\varepsilon}\right), \quad C_{ij} = c(x_i, y_j)$$

Solution form:

$$\pi^\varepsilon_{ij} = u_i K_{ij} v_j$$

Sinkhorn Iterations

Iterative updates:

$$u^{(k+1)} = \mu/(Kv^{(k)})$$
$$v^{(k+1)} = \nu/(K^\top u^{(k+1)})$$

Initialization: $u = \mathbf{1}, v = \mathbf{1}$

Convergence:

- Converges geometrically.
- Fast even for large-scale problems.
- Stabilized versions available for small ε.

Output: $\pi^\varepsilon = \mathrm{diag}(u) K \,\mathrm{diag}(v)$

Python Pseudocode: Sinkhorn Algorithm

[language=Python] def sinkhorn(mu, nu, C, epsilon=0.01, max$_i ter = 1000, tol = 1e - 9) : K = np.exp(-C/epsilon) u = np.ones_l ike(mu) v = np.ones_l ike(nu) for_i nrange(max_i ter) : u_p rev = u.copy() u = mu/(K@v) v = nu/(K.T@u) ifnp.linalg.norm(u - u_p rev, 1) < tol : breakreturnnp.diag(u)@K@np.diag(v)$

Sinkhorn: Summary and Benefits

Why Sinkhorn?

- Reduces OT to scalable matrix scaling.
- Smooth objective allows automatic differentiation.
- Central to modern machine learning: generative models, domain adaptation, GANs.

Sinkhorn vs. Classic OT:

- Classical OT is slow (LP solver).
- Sinkhorn is fast, parallelizable, and differentiable.

Next: Applications of geometry-aware transport in statistics and inference.

Motivation and Aim

Goal: Provide a unified, geometric theory for financial product design that applies to both **hedging** and **investment**.

- Built on *information derivatives*, encoding beliefs via probability distributions.
- Employs tools from **information geometry** to understand market scenarios, investor behavior, and product risks.
- Connects utility theory, Bayesian inference, and KL divergence to finance.

Unified Probabilistic Framework

Financial decisions involve beliefs represented as probability distributions:

- Market-implied (prior): $m(x)$
- Investor-believed (posterior): $b(x)$
- Scenario: perturbations or alternatives

Key mechanism: Likelihood product:

$$b(x) = f(x)m(x)$$

where $f(x)$ is the *likelihood function*, interpreted as a **payoff structure**.

Introduction Summary

- Financial products encode *views* via payoffs.
- Hedging and investment both reduce to belief-based optimization.
- Structure of beliefs and products is inherently **geometric**.
- Sets stage for risk to be understood as a form of **return differential** and **divergence**.

Multiple Rationality (Sec 2.1)

Observation: Human behavior is driven by multiple, goal-specific rationalities.

- Not globally irrational — rather, **multi-rational**.
- Each product targets a specific function or goal.
- This justifies using *expected utility theory* at the product level.

Example: A person may simultaneously seek safety (via insurance) and growth (via investment).

Financial Product as a Payoff Function

Definition: A financial product is defined by a payoff function $F(x) \geq 0$

- Allows scaling: $F(x) \sim \lambda F(x)$ (notional multiplier)
- Product is an asset: non-negative payoff
- Viewed as response to an optimization problem

Aim: Build a **scientific theory** of product design — i.e., consistent with data and human reasoning.

Bayesian Foundations: Likelihood Product

Bayes' Rule: Posterior belief from market and research:

$$b(x) = f(x)m(x)$$

- $m(x)$: Market-implied (prior)
- $b(x)$: Investor belief (posterior)
- $f(x)$: Likelihood = **investment product** encoding research

This defines the **likelihood product**.

Investor Equivalence Principle

For any payoff $F(x)$:

$$\omega_F(x) = F(x)m(x)$$

- ω_F is an implied distribution: a *view*.
- Can always find an equivalent likelihood investor.
- Normalizing ω_F gives a probability distribution.

Implication: Focus on likelihood investors captures realistic product behavior.

General Rational Product: Utility Maximization

Investor optimization problem:

$$\max_F \int b(x)U(F(x))dx \quad \text{s.t.} \quad \int F(x)m(x)dx = 1$$

Solution: The **payoff elasticity equation**:

$$\frac{d \ln F}{d \ln f} = \frac{1}{R(x)}$$

where $R(x) = -\frac{xU''(x)}{U'(x)}$ is the Arrow-Pratt relative risk aversion.

Utility Functions and Risk Aversion

Utility encodes preferences over uncertain outcomes.

- Risk-neutral: $U(x) = x \Rightarrow \mathrm{R(x)} = 0 \mathit{CRRA} : \mathrm{U(x)} = \mathrm{x}^{1-\gamma} \frac{}{1-\gamma} \Rightarrow \mathrm{R(x)} = \gamma$

- Log utility: $U(x) = \log x \Rightarrow \mathrm{R(x)} = 1 \mathit{Exponential} : \mathrm{U(x)} = 1 - \mathrm{e}^{-x} \Rightarrow \mathrm{R(x)} = \mathrm{x}$ *Higher* $R(x)$ *implies more aversion to risk.*

Key Takeaways from Section 2

- Products are best understood through the **likelihoods** they encode.
- Any payoff $F(x)$ implies a view ω_F.
- Rational investors choose F by maximizing expected utility.
- The elasticity equation relates payoff shape to belief and risk preferences.

This prepares the ground for Section 3: understanding **risk as spread in returns**.

Section 3: Risks as Returns

Key idea: Price sensitivity (risk) can be expressed as a **return spread**.

- Risk is not just variance — it is the expected log-return differential.
- Uses *likelihood products* to measure exposure to scenarios.
- Paves the way to defining risk geometrically.

Portfolio Payoff and Price

Let $\Pi(x)$ be a portfolio payoff function. Market-implied price:

$$\text{Price}[\Pi] = \int \Pi(x) m(x) dx$$

Small perturbation in market belief:

$$b(x) = m_{\omega+\varepsilon}(x), \quad f(x) = \frac{b(x)}{m(x)}$$

Perturbed price sensitivity:

$$\frac{d}{d\varepsilon}\text{Price}[\Pi] = \int \Pi(x) \frac{\partial m(x)}{\partial \varepsilon} dx$$

Exponential Score Product

In the risk-neutral limit (small ε, $R \to 0$), the optimal product becomes:

$$F_0(x) = \frac{e^{\mathrm{Score}(x)}}{\mathbb{E}_m[e^{\mathrm{Score}}]}, \quad \mathrm{Score}(x) = \frac{\partial}{\partial \varepsilon} \ln m(x)$$

Then the portfolio-implied view:

$$\omega_\Pi(x) = \frac{\Pi(x)}{\mathrm{Price}[\Pi]} m(x)$$

This is a probability distribution implied by holding Π.

Defining Risk via Returns

Specific risk (risk per unit price) of Π with respect to product S:

$$\mathrm{Risk}_S[\Pi] = \mathrm{Price}[\Pi] \cdot (\mathbb{E}_{\omega_\Pi}[\ln S] - \mathbb{E}_m[\ln S])$$

- Measures **difference in expected log-returns** under the two distributions.
- For $S = F_0$, this reduces to the **standard sensitivity**.
- Crucially: S captures the **scenario** under which we assess risk.

Numerical Example: Risk as Return Spread

Suppose:

- Two outcomes: $x = 0, 1$
- Market: $m = [0.6, 0.4]$
- Investor: $b = [0.5, 0.5]$
- Then $f(x) = \left[\frac{5}{6}, \frac{5}{4}\right]$, and $F_0 \approx [1.114, 0.734]$

Compute:

$$\begin{aligned}
\mathbb{E}_b[\ln F_0] &= 0.5\ln(1.114) + 0.5\ln(0.734) \approx -0.101 \\
\mathbb{E}_m[\ln F_0] &= 0.6\ln(1.114) + 0.4\ln(0.734) \approx -0.059 \\
\Rightarrow \mathrm{Risk}_{F_0}[\Pi] &= -0.042
\end{aligned}$$

Interpretation: Portfolio has negative exposure to the scenario.

Section 3 Summary

- Risk $=$ **return differential** between investor and market expectations.
- Specific risk: $\mathrm{Risk}_S[\Pi] = \mathbb{E}_{\omega_\Pi}[\ln S] - \mathbb{E}_m[\ln S]$
- F_0 is a canonical risk scenario product for infinitesimal perturbations.
- This formulation sets up Section 4: **Information geometry of risk**.

Objective: Understand risk geometrically using KL divergence and dual geodesics.

- Risk defined as a configuration of three distributions:
- ω_Π: portfolio-implied belief
- m: market-implied belief
- ω_S: risk scenario

KL Divergence and Risk Geometry

Definition: Kullback-Leibler divergence

$$D(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

Geometric identity for risk:

$$\text{Risk}_S[\Pi] = D(\omega_\Pi \| m) + D(m \| \omega_S) - D(\omega_\Pi \| \omega_S)$$

This forms a **triangle** in information space.

Geometric Interpretation

Three distributions form a triangle:

- ω_{Π}: expresses investor view
- m: market consensus
- $\omega_S = Sm$: scenario under risk product

Risk sign determined by angle at m:

- Acute angle $\Rightarrow$ positive risk
- Right angle $\Rightarrow$ zero risk
- Obtuse angle $\Rightarrow$ negative risk

Mixture and Exponential Geodesics

Define two interpolations:

$$p_{\text{mix}}(x,t) = (1-t)m(x) + t\,\omega_{\Pi}(x)$$

$$p_{\text{exp}}(x,t) = \frac{m(x)^{1-t}\omega_S(x)^t}{Z(t)}$$

Interpretation:

- p_{mix}: movement toward liquidating portfolio (m-geodesic)
- p_{exp}: movement toward risk scenario (e-geodesic)

Scalar Product of Tangents

Define geodesic tangent vectors:

$$\frac{d}{dt} p_{\text{mix}}(x,t)\bigg|_{t=0} = \omega_\Pi(x) - m(x)$$

$$\frac{d}{dt} \ln p_{\text{exp}}(x,t)\bigg|_{t=0} = \ln \omega_S(x) - \ln m(x)$$

Their inner product yields specific risk:

$$\langle \omega_\Pi - m, \ln \omega_S - \ln m \rangle = \text{Risk}_S[\Pi]$$

Iso-Risk Foliation

Iso-risk surfaces: Distributions with constant Risk_S.

- They form **m-flats**, i.e., flat under the m-geometry.
- Orthogonal to the e-geodesic connecting m and ω_S.
- Adding S to Π moves us along the e-geodesic toward ω_S.

Illustration: Geometric Triangle of Risk

triangle_geometry.png

Section 4 Summary

- Risk has a precise geometric structure via KL divergence.
- Mixture and exponential geodesics describe portfolio and scenario dynamics.
- Iso-risk surfaces and their orthogonality explain hedging logic.
- Sets stage for optimal product design in Section 5.

Section 5: Hedging with Information Derivatives

Goal: Design hedging products using geometric intuition from KL divergence.

- Use products to eliminate (or control) specific risks.
- Leverage the geometry: move portfolio views to iso-risk surfaces.
- Optimize cost and expressiveness of hedges.

Hedging via e-Geodesics

Start from ω_Π with $\text{Risk}_S[\Pi] \neq 0$.

- Add exposure to S to move along the e-geodesic toward ω_S.
- e-geodesic intersects iso-risk surface (zero-risk manifold).
- Simple hedge: find t such that $\text{Risk}_S[\omega_t] = 0$

Equation:

$$\omega_t(x) = \frac{m(x)S(x)^t}{\int m(y)S(y)^t dy}$$

Monotonicity and Search for Hedge

Risk exposure evolves along e-geodesic:

$$\mathrm{Risk}_S[\omega_t] = \int_0^t \mathrm{Var}_{\omega_s}[\ln S]ds$$

Implication:

- $\mathrm{Risk}_S[\omega_t]$ increases monotonically in t
- Use simple 1D search to find exact hedge level t.

Cost-Optimal Hedge (c-Projection)

Goal: Minimize trading cost while neutralizing risk.

- Cost is modeled as pointwise function: $C(x, y)$
- Solve for adjusted payoff $\Pi_{\rightarrow}(x) = \Pi(x) + \delta(x)$ minimizing:

$$\int C(x, \delta(x))dx \quad \text{s.t. } \mathrm{Risk}_S[\Pi_{\rightarrow}] = 0$$

Solution: $\delta(x)$ is a monotonic function of $\ln S(x)$.

Pure Hedging Products

Definition: A hedge that expresses *minimal view* while achieving risk control.

- Optimal solution: closest distribution to m with given risk exposure.
- Formally:

$$\omega_H = \arg\min_{\omega} D(\omega \| m) \quad \text{s.t. } \mathrm{Risk}_S[\omega] = r$$

Geometric solution: ω_H lies on the e-geodesic from m to ω_S.

Generalized Divergences: $\phi - DivergenceHedging$

Replace KL divergence with a more general divergence:

$$D_\phi(\omega \| m) = \int m(x)\phi\left(\frac{\omega(x)}{m(x)}\right) dx$$

- Minimization over D_ϕ still yields monotonic hedge structures.
- Optimal hedge: $H(x) = M_\phi(S(x))$ for some increasing/decreasing map.

Hedge-Investment Duality

Key observation: An optimal hedge is also a rational investment.

- Investment maximizes expected utility:

$$F = \arg\max \int b(x)U(F(x))dx \quad \text{s.t. } \text{Price}[F] = 1$$

- Same F can also be derived as a pure hedge w.r.t. S with divergence D_ϕ

Conclusion: Investments and hedges are geometrically dual objects.

Risk Recycling

Idea: Sell a hedge as an investment product to a client with matching view.

- For $\text{Risk}_S[A] > 0$, design hedge H s.t. $\text{Risk}_S[H] = \text{Risk}_S[A]$.
- Find a belief $b = S \cdot m$ and utility U for which H solves

$$F = \arg\max \int b(x)U(F(x))dx$$

This enables: *Repackaging risk as transparent rational investments.*

Partial Hedging: Constrained Utility Optimization

Want: Maximize utility with fixed risk exposure:

$$\max_F \int b(x)U(F(x))dx \quad \text{s.t. } \text{Risk}_S[F] = r$$

Result: Modified elasticity equation:

$$\frac{d\ln F}{d\ln f} = \frac{1}{R} \cdot \left(1 - \frac{d\ln(1+\lambda \ln S)}{d\ln f}\right)^{-1}$$

This allows initial delta to be set, generalizing swap-format structures.

Section 5 Summary

- Hedging = moving ω_{Π} to iso-risk surfaces via product design.
- Pure hedges = minimal divergence adjustments from m.
- Cost-optimal hedging: c-projection under trading cost.
- Hedges and investments form a duality; risks can be recycled transparently.
- Partial hedging possible via constrained utility maximization.