# An Intrinsic Theory of Information Acquisition: Application to Dynamic Portfolio Choice, Asset Pricing and Information Recovery

Raymond C. W. Leung[*]

February 2018

**Abstract**

We introduce an intrinsic geometric framework to model the optimal path of information acquisition. Agents' utility maximizing actions are naturally dependent on the path on which they acquire information. By an intrinsic framework we mean to focus purely on the geometry of information acquisition that is invariant to arbitrary parameterization and information measurement by an outside econometrician. Intrinsic curvature of the information manifold is the key driver of our framework: curvature relates to the marginal cost of information acquisition that is averaged across all possible information acquisition directions. In a finance application, our geometric framework leads to canonical definitions of velocity, acceleration, curvature and torsion of portfolio holdings and equilibrium returns of risky assets. To demonstrate the empirical applicability potential of our framework, we show that "return absolute curvature" is a significant explanatory variable of VIX. Our geometric framework also provides a method for information recovery: upon observing a trajectory of portfolio holdings by an econometrician, we can geometrically infer the mean-variance information used by the portfolio manager. Finally, our framework is general: essentially all expected utility maximization problems based on parameterized random variables are applicable.

**JEL classification: D83, G11, G12**

**Keywords:** information acquisition, portfolio choice, asset pricing, information geometry, Riemannian geometry

[*] Department of Finance, Cheung Kong Graduate School of Business, email: raymondleung@ckgsb.edu.cn. I thank Farshad Haghpanah, Zhongzhi Song, and Yu-Man Tam for helpful and insightful discussions.

# 1 Introduction

It is indisputable that endogenous acquisition of information is the key to understanding how economic agents make decisions. The extant literature of information economics have focused on how equilibria are influenced by the presence of informed and uninformed agents. While we have learned much from this literature on *what* happens when economic agents have heterogeneous information sets, this literature has remained silent on *how* an agent obtains their own piece of knowledge in the first place. To take a concrete example in finance, suppose an investor is endowed with an initial belief of the returns of some risky assets. If it is costly to learn information above and beyond his initial belief, it is natural to ask which direction of information should he acquire first. When should the investor learn more about the mean of the return, the variance, or both? If so, at what velocity and at what acceleration of learning?

We present in this paper an *intrinsic framework for information acquisition*, and show its applicability to classic finance problems. The framework is *intrinsic*, meaning we solely concentrate on the geometric properties of information acquisition. In contrast, an *extrinsic* theory of information acquisition will heavily depend on arbitrary parameterizations and measurements of an agent's information set by a theory-modeller or an econometrician. In effect, this arbitrariness in describing an extrinsic framework for information acquisition imply that different theory models are largely incompatible and incomparable to each other.

Our intrinsic framework is based on the well-established field of *information geometry* in probability and statistics theory. To the best of my knowledge, information geometric methods have not been applied to theoretical economic and finance problems. The key to an intrinsic theory of information acquisition is recognize that: (P1) states of knowledge can be represented by probability distributions; (P2) knowledge should be intrinsically and not extrinsically measured; and (P3) economic agents' expected utility maximizing action choices should be based on intrinsically described information. We call these three points as *principles for an intrinsic theory of information acquisition.* Principles (P1) and (P2) are largely resolved by information geometric methods. Our consideration for endogenous optimization problems in (P3) differentiate our paper from just rehashing the existing information geometry literature.

We canonically use the *distance* between knowledge (i.e. probability distributions) to measure the cost of information acquisition from one point to another point of knowledge. In adherence to principle (P2), the cost function is canonical and intrinsic: the choice of uncertainty form in the economy *fixes* the distance measure. In this sense, the functional form of the cost of information acquisition is not completely arbitrary, and thus disciplines the theory-modeller and econometrician. A key contribution of our framework is that we can intrinsically describe the *trajectory* of information acquisition at the *minimal* cost. As a result, we can concretely answer our finance motivated question in the opening paragraph and describe the velocity, speed and acceleration of knowledge acquisition of risky asset returns. We show that the key and essentially only driver in our intrinsic geometric framework is the *curvature* of the manifold of information. The cost of information acquisition in this geometric framework is non-constant. In particular, the marginal cost of acquiring one direction of knowledge versus the marginal cost of another direction actually depends on the knowledge state. In that finance motivated example, this means depending on the investor's position of his initial endowed belief of the risky asset return, he may find it more or less costly to first acquire knowledge in the mean direction versus the variance direction. This causes the optimal information acquisition trajectory to "curve". Overall, this "curvature" of the manifold of information summarizes the average marginal cost of information acquisition.

Finally, the application of our geometric information acquisition framework to a classic portfolio choice and asset pricing problem will yield new additional insights. If our investor has CARA preferences and is myopic, then we can discuss a geometry of the optimal portfolio allocations along an optimal information acquisition trajectory. Specifically, we can define notions like portfolio trajectory velocity, acceleration, curvature and torsion. These notions endogenously arise out of the curvature in the manifold of information. In equilibrium, we further show the instantaneous return dynamics also enjoy several geometric notions like return velocity, acceleration and curvature. As a demonstration for the empirical applicability, we empirically show that "return absolute curvature" is a significant explanatory variable for VIX. One hopes that the introduction of these geometric concepts to describing empirical return behavior might lead to new additional insights in future research.

# 2 Motivating information geometry

While the field of *information geometry* is well established in the statistical and probability theory literature [1] , its applications are not well known in economic theory [2] . To motivate why it is useful to consider information geometric techniques, let's first briefly review the contemporary and predominant methods of modelling information in the economics literature. Typically in these models, there a signal $U$ of an economic object of interest, but an agent can only observe a noisy version $S = U + \epsilon$, where $\epsilon$ is a random noise term with mean zero that prevents the agent from perfect observation of the signal $V$. It is also typically assumed that the agent can pay some cost to increase the precision $1/\text{Var}(\epsilon)$ of the noise term. This form of modelling information is widely used in applied theories from contracting, accounting, trade, and finance. A likely reason for the widespread adoption of this modelling form is its simplicity.

Despite the widespread and successful adoption of this form of information modelling, it nonetheless suffers a deficit: what concretely is a "noisy observation"? More importantly, can this modelling approach say anything concrete about the "type" and "direction" of information an agent should acquire? For instance, should an agent learn more about the "mean", "volatility" or other moments of the information? How quickly should the agent learn about these moments and in which direction? How does the direction and speed of learning depend on his prior information? What should be the trajectory of learning? The current information modelling paradigm is completely silent to these arguably important and concrete questions.

We will motivate why information geometric tools are natural in answering important questions in information acquisition in three different ways. All three ways revolve around how one views the "distance" between informations.

## 2.1 What is knowledge?

We must first explicitly discuss what we mean by *information* or *knowledge*; throughout this paper, we will interchange using these two words for the same meaning. Simply put, *a*

[1] See the textbook treatments by Amari and Nagaoka (2007), Calin and Udriste (2014), Amari (2016), and Ay et al. (2017).

[2] Geometric methods have been selectively applied in econometrics; see Marriott and Salmon (2000) for a textbook treatment. Debreu (1972) used differential geometric methods to investigate preference relations.

*probability assignment of an event describes a certain state of knowledge.* This principle has had a long history in probability, statistics and even in economics. Bernoulli (1713) calls it the *Principle of Insufficient Reason* in *Ars Conjectandi.* The esteemed economist John Maynard Keynes calls it the *Principle of Insufficient Reason* in his *A Treaties of Probability* (Keynes (1921)). Jaynes (1978) recounts a colorful history on this principle of viewing probability as representing the state of knowledge. However, this way of viewing probabilities took a sharp turn by the late 18th to early 19th century. Indeed, Jaynes (1978) writes:

> "This counter-stream of thought, however, rejected the notion of probability as describing a state of knowledge, and insisted by 'probability' one must mean only 'frequency in a random experiment'. For a time this viewpoint dominated the field so completely that those who were students in the period 1930-1960 were hardly aware that any other conception had ever existed".

Sampling or frequentist interpretation of probability took over most of the 19th century.

However by the 1950's, a "Bayesian revolution" in statistics and probability came about. In a nutshell, if one accepts the Bayes rule, then one must accept that probability represents knowledge, and not the frequency of which an event happens. From there, ideas of *entropy* — which we will heavily use in this paper — flourished. Again, Jaynes (1978) has a wonderful historical account of these philosophical developments:

> "[T]o a person who has been trained to think probability only in the sense of frequency in a random experiment (as was surely the case for anyone educated at M.I.T. in the 1930's!), the idea that a probability distribution represents a mere state of knowledge is strictly taboo."

In this paper, we will axiomatically accept that each random variable, and their associated probability distribution, represents the state of knowledge of an agent. Under this principle, knowledge acquisition then can be mathematically and economically represented as "moving" from one random variable (representing one probability distribution) to another random variable (representing another probability distribution). We devote a good portion of the paper to discuss what does it mean and how does one "move" from one knowledge point to another. These ideas will be made precise in Section A.1. In order to discuss how to "move" knowledge, we must discuss how to measure the distance of knowledge.

## 2.2 Rational inattention

The *rational inattention* literature provides a natural motivation for discussing the distance between knowledge. The key idea in this literature is an agent's "limited capacity for processing information" (Sims (2003)). In particular, suppose $X_{p_0}$ be a parameterized random variable that represents the "truth information" and let $X_p$ be some "action information" the agent can acquire. Here we deliberately use the alphabet letters, like $p$, to denote a parametrization, rather than the perhaps more conventional notations that use Greek letters; this is to harmonize with the geometry based notations for the rest of the paper.

Since the agent cannot immediately process all information, $I(X_{p_0}; X_p)$ (the *mutual information*) measures the information gain for the agent for acquiring the truth $X_{p_0}$ relative to his current information $X_p$. Thus effectively, the rational inattention literature takes the stance that the distance between random variables — as measured by mutual information — measures the cost of acquiring information, and this cost arises precisely because of the agent's limited information processing capacity.

For our purposes, it is technically more convenient to not work with mutual information $I(X_{p_0}; X_p)$, but rather with *relatively entropy* (or also frequently known as *Kullback-Leibler divergence (KL divergence)*) $D_{\mathrm{KL}}(X_p || X_{p_0})$. Sims (2003) draws out the connection between his mutual information based framework to relative entropy. Mutual information and relative entropy are related as $I(X_{p_0}; X_p) = \mathbb{E}_{X_{p_0}} D_{\mathrm{KL}}[(X_p | X_{p_0}) || X_p]$, where $X|Y$ is the conditional distribution of $X$ given $Y$. Like mutual information, the relative entropy can also be viewed as a form of "distance" between two random variables. With this interpretation, we can think of $D_{\mathrm{KL}}(X_p || X_{p_0})$ as the distance between information from an agent's current information $X_\theta$ relative to the truth $X_{\theta_0}$, and this distance again measures the cost of the agent's limited information processing capacity.

Most crucial for our purpose is that relative entropy is intimately related to the *Fisher information matrix.* Rao (1945) made the pioneering connection between the distance of information by using the Fisher information matrix as a metric. The Fisher information matrix of a parameterized probability distribution $f(x; p)$, where $p = (p^1, \ldots, p^m)^\top$ has its $(j, k)$-th entry as,

$$g_{jk}(p) := \mathbb{E}\left[\left(\frac{\partial}{\partial p^j} \log f(X; p)\right)\left(\frac{\partial}{\partial p^k} \log f(X; p)\right)\right]. \tag{2.1}$$

It is easily shown that $[g_{jk}(p)]$ is an $m \times m$ positive semidefinite symmetric matrix. In the rest of our paper, we will only consider distributions where the Fisher information matrix is actually positive definite.

**Lemma 2.1** (KL divergence and Fisher information metric, Jeffreys (1946))**.** *Let $f(\cdot; p)$ denote the probability density function of the random variable $X_p$ and let's consider $p$ close to $p_0$. For $\theta \mapsto f(\cdot; p)$ sufficiently smooth, we have that*

$$D_{KL}(X_p || X_{p_0}) \approx \frac{1}{2} \sum_{j,k} \Delta p^j \Delta p^k g_{jk}(p_0),$$

*where $\Delta p^j := (p - p_0)^j$ is the $j$th component of the vector $p - p_0$.*

The key observation from Lemma 2.1 is that when we consider two infinitely close probability distributions, relative entropy and the Fisher information matrix capture the same notion of distance between random variables. As we shall see, there are significant economic insights to be gained with using the Fisher information matrix as a measure of distance between random variables rather than the KL divergence or mutual information. Ly et al. (2017) provide an overview of the Fisher information matrix, and draws the connection between this matrix to mathematical psychology and cognitive modeling.

## 2.3 Bayesian news acquisition

As we have seen above, an agent's limited information processing capacity can motivate why it is important to study the distance between information (or random variables). Here, we provide an alternative justification from an information acquisition perspective. For concreteness, let's specify our discussion to an univariate Gaussian distribution, although our discussion works for any general multivariate parameterized distributions

Suppose $Z(T) \sim \mathcal{N}(\mu(T), \sigma(T))$ represents the true distribution of some piece of information that is only revealed at some terminal time $T$. The agent is endowed with some "initial knowledge" $Z(0) \sim \mathcal{N}(\mu(0), \sigma(0))$ at time $t = 0$. For instance, $Z(T)$ is true return distribution of a risky asset, and $Z(0)$ is the agent's initial knowledge of this true return. The objective of the agent is to acquire the true distribution $Z(T)$ from his endowed knowledge $Z(0)$ through an endogenous "news trajectory". News does not arrive instantaneously to the agent. In particular, there is no one single piece of news for which the agent can "jump learn" from his initial knowledge $Z(0)$ to the terminal truth $Z(T)$.

As a result, the agent must *endogenously* choose the type and direction of news acquisition. In particular at each time $t$, the agent has some prior knowledge $Z(t) \sim \mathcal{N}(\mu(t), \sigma(t))$. Using the Bayes rule, the agent selects and acquires a piece of news at time $t$ such that his posterior knowledge for the next period $t + \mathrm{d}t$ becomes $Z(t + \mathrm{d}t) \sim \mathcal{N}(\mu(t + \mathrm{d}t), \sigma(t + \mathrm{d}t))$. Once the agent reaches to time $t + dt$, the agent uses $Z(t + \mathrm{d}t)$ as his prior (which was previously the time $t + \mathrm{d}t$ posterior) to update to his next posterior $Z(t + 2\mathrm{d}t)$ at time $t + \mathrm{d}t$. The agent iterates this information acquisition strategy from $t = 0$ to $t = T$. See Figure 1 for an illustration.

Next, we describe how the agent optimally picks his information trajectory. From a purely statistical standpoint, the agent wants to pick the sequence news such that the "distance" between the random variables $Z(t)$ and $Z(t + \mathrm{d}t)$ is small, and all the while ensuring he can reach from $Z(0)$ to $Z(T)$. One can view this "distance" as the cost of acquiring news to update the agent from an old knowledge $Z(t)$ to the new knowledge $Z(t + \mathrm{d}t)$. Furthermore, Section 2.2 provides an alternative microfoundation for this "distance": if the agent has a limited capacity for processing information, then the agent will pick information such that the "mental processing cost" (namely as measured by relative entropy) between $Z(t)$ and $Z(t + \mathrm{d}t)$ is small.

The following result will be an easy corollary from one of the results of our paper. We state the result in its full generality, meaning that $Z(t)$ in the statement can be multivariate, non-Gaussian, discretely or continuously distributed.

**Lemma 2.2** (Optimal Bayesian news acquisition trajectory)**.** *Consider the framework described in Section 2.3, and let $Z(0)$ be an agent's initial knowledge of an economic variable, and $Z(T)$ represent its terminal truth knowledge. Then:*

*(i) For any information trajectory $\{Z(t)\}_{t\in[0,T]}$, there exists a Bayesian likelihood between all times $t$ and $t + \mathrm{d}t$ such that $Z(t)$ is the prior and $Z(t + \mathrm{d}t)$ is the posterior.*

*(ii) Suppose $Z(0), Z(T)$ belong to the same parameterized distribution family, and whose distribution support is* not *dependent on the parameters. Then there always exist an information trajectory (possibly non-unique) $\{Z(t)\}_{t\in[0,T]}$ that starts at $Z(0)$, ends at $Z(T)$, and that minimizes the relative entropy between the $Z(0)$ and $Z(T)$.*

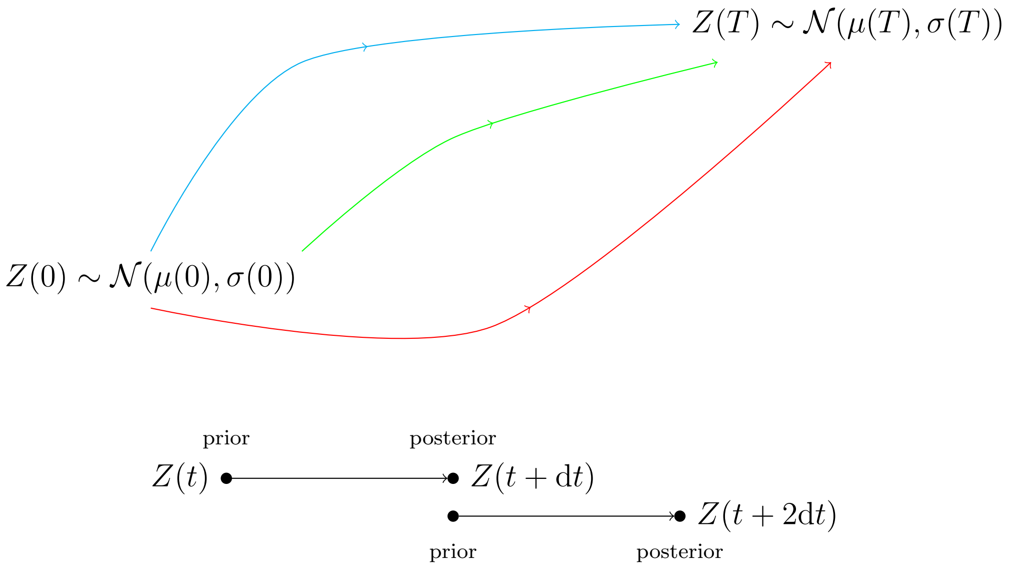
$Z(T) \sim \mathcal{N}(\mu(T), \sigma(T))$
$Z(0) \sim \mathcal{N}(\mu(0), \sigma(0))$
prior
posterior
$Z(t)$
$Z(t + \mathrm{d}t)$
$Z(t + 2\mathrm{d}t)$
prior
posterior

**Figure 1:** Bayesian news acquisition trajectory

*Proof of Lemma 2.2.* Part (ii) will be an immediate result of Proposition 5.1. □

The proof of statement (a) is immediate from Bayes theorem. Statement (a) simply says if there is a sequence of random variables $\{Z(t)\}$ where each $Z(t)$ are independent and belong to the same distribution family, then we can always find a Bayesian information acquisition interpretation to link $Z(t)$ and $Z(t + \mathrm{d}t)$. Statement (b) forms the core microfoundations of our *intrinsic information economic* framework that we will formalize in Section A. It tells us that given only some initial knowledge and some terminal knowledge, it is *always* possible to find an information acquisition trajectory $\{Z(t)\}$ that connects them. Moreover, this information acquisition trajectory is optimal in the sense that it minimizes the "distance" between them. One interpretation is that there is always a cheapest way to acquire news. Alternatively, using the rational inattention interpretation, an agent can always find a way to acquire information to minimize his mental information processing cost. As we shall see in more detail in Proposition 5.1, (b) is a consequence of a very deep result and should not be deemed as "obvious".

## 2.4 Intrinsic versus extrinsic perspective of information

Either through a rational inattention or an Bayesian news acquisition perspective, we hope it is clear by now that the notion of distance between random variables is critical to understanding endogenous information acquisition. However, we must be careful in understanding that information must be "intrinsic" and not "extrinsic" in the following sense. We will again use the Gaussian distribution to illustrate the idea. As it is well known, the Gaussian distribution can be characterized by two parameters. One possible parameterization is via its mean and standard deviation. Suppose now we have two Gaussian random variables $X_1$ (mean 0 and standard deviation 2) and $X_2$ (mean 0 and standard deviation 4). Nonetheless, this mean and standard deviation parameterization form, however conventional, is still arbitrary. We could have equally described the two random variables as $Y_1$ (mean 0 and precision 1/2) and $Y_2$ (mean 0 and precision 1/4). If we "naively" use the $\mathbb{R}^2$ Euclidean distance to measure the parameterization distance between $X_1$ and $X_2$, we would have $||(0,2)-(0,4)||_{\mathbb{R}^2} = 2$. However, if we were to use the same $\mathbb{R}^2$ Euclidean distance to measure the parameterization distance between $Y_1$ and $Y_2$, we would have $||(0,1/2)-(0,1/4)||_{\mathbb{R}^2} = 1/4$. This is awkward — two equivalent methods of describing the random variables yield two different distance values. The notion of distance should be invariant (in some important sense to be discussed in Section A) to arbitrary parameterizations. In more colorful but suggestive language, the value of information should be independent of whether it is written in English or in Japanese!

The following analogy will be useful to keep in mind starting from Section A below when we formally introduce geometric aspects to our framework. When a bird flies by in the physical world, it simply does *not* care whether an observing physicist measures its flight trajectory in meters or in feet. A good physical theory for the trajectory should be *intrinsic* to the bird, and not depend on subjective *extrinsic* choices of an observer. By analogy, how an agent acquires knowledge in his psychological mind should not depend on how an econometrician measures it.

In all, an "extrinsic" notion of information will depend significantly on the arbitrary parameterization of an outside observer, while an "intrinsic" notion of information should be free of this arbitrariness. As we shall see, the appropriate intrinsic measure of information will be given by the Fisher information metric of (2.1).

# 3 Intrinsic theory of information acquisition

Regardless of the perspectives one adopts from Section 2, we enforce these principles that should hold for an *intrinsic theory of information acquisition*:

**Principles for an intrinsic theory of information acquisition** *We postulate that an intrinsic theory for information acquisition should simultaneously satisfy the following three principles:*

*(P1) Probability distributions represent the state of knowledge of an agent.*

*(P2) Information should be intrinsically and not extrinsically measured;*

*(P3) Any utility maximizing action by an agent needs to be based on his intrinsic knowledge.*

*Information geometry* offers a way to satisfy both principles (P1) and (P2), and provide a natural distance between information. This natural distance between information will be our cost function for information acquisition. Principle (P3) is the key that elevates and differentiates our paper from existing information geometry papers.

The literature on information geometry is vast, and is built upon the theories of *differential geometry* and *Riemannian geometry*. Roughly speaking, *information geometry* puts a geometric manifold structure on the space of probability distributions. Differential geometry is the study of calculus on manifolds. Riemannian geometry introduces a metric structure on manifolds on top of a calculus structure.[3] .

## 3.1 Finance and economic applications roadmap

We are most interested in applying information geometry to concrete economic applications. Indeed, without these critical applications, our discussions here would simply rehash the literature of information geometry. In this paper, we focus on three sequentially related finance questions:

[3] Differential and Riemannian geometry found their most substantial application in Einstein's *general relativity* in physics. As history goes, apparently mathematician Marcel Grossman was instrumental in influencing and convincing Einstein the importance of non-Euclidean geometry in developing the theory of general relativity.

1. **Dynamic portfolio allocation (Section 6).** Suppose an investor is endowed with some initial knowledge of return distribution of some risky assets, and chooses an optimal information acquisition trajectory to reach the true distribution. Each point along the optimal trajectory represents the investor's best available or closest knowledge to the truth. If the investor myopically allocates portfolio choices between these risky assets to reflect based on his best available knowledge, what is the *velocity* and *acceleration* of the investor's portfolio path?

2. **General equilibrium asset pricing (Section 7).** Information in the economy is highly heterogeneous. Some investors may be completely uninformed and need to acquire information on all aspects of the return distribution. Yet some investors may be partially informed and only needs to learn a certain aspect. In equilibrium, how do *asset returns*, and their *time variability* depend on the myriad of heterogeneous information acquisition trajectories?

3. **Information recovery from observed portfolio choices (Section 8).** Unlike Questions 1 and 2 where we start from an agent's information acquisition trajectory and ask what are the portfolio choice and asset pricing implications, we now ask the inverse question: Suppose an econometrician observes only the portfolio allocations of an investor over time. Can the econometrician *recover* the information the investor used to base his investment decision? Can the econometrician *distinguish* between an informed investor (who acquires information to reach the truth) versus an uninformed investor (who arbitrarily uses information)?

Actually, more is true than the above three finance specific applications. These three applications essentially "prove by concept" (of course we will rigorously prove) the following result:

4. **Practically all expected utility maximization problems can inherit this intrinsic information acquisition framework (Section 9).** Beyond finance specific applications, our intrinsic information acquisition framework enjoy a sweeping generalization to practically all problems that involve expected utility maximization. In fact, our framework provides a "poor man's way" of injecting information acquisition dynamics to one-period expected utility maximization problems.

One of the key result that we need from information geometry is the *geodesic* equation in Section A.3. In an information geometry context, geodesics are precisely those optimal information acquisition trajectories of Lemma 2.2. One needs actually substantial amount of background concepts to describe the geodesic equation, and we devote Sections A for the background materials at a strictly intuitive and graphical level.

For our economic applications, we are only interested in the geodesic equations of Gaussian distributions, and these results are reported in Section 5. If the reader is only interested in applying the geodesic equation, and omit on a first reading the justification of the equation, the reader may skip all of Section A and proceed directly to Section 5. The reader must then, *prima facie*, accept that geodesics are the correct intrinsic way to describe the shortest trajectory between some initial knowledge and terminal truth.

## 3.2 Notations and Einstein's summation convention

We adopt notations from the differential geometry literature. We write the components of a vector $v = (v^1, \ldots, v^m)$ with an upper index. We also adopt *Einstein's summation convention*: an index variable that appears twice in a term implies summation of that term over all its index values. For example, in $\mathbb{R}^m$ and letting $E_i = (0, \ldots, 1, \ldots, 0)$ denote the standard basis, a vector $v$ will be written as,

$$v = v^i E_i := \sum_{i=1}^{m} v^i E_i.$$

An $n \times m$ matrix $A$ has element $A^i_j$ in its $i$th row, and $j$th column. In particular this means the matrix multiplication, when $A$ is $n \times m$ and $v$ is $m \times 1$, $y = Av$ has in the $k$th component of $y$,

$$y^k = \sum_{j=1}^{m} A^k_j v^j = A^k_j v^j. \tag{3.1}$$

# 4 Overview of information geometry, differential geometry, and Riemannian geometry

This paper does *not* intend to be a treatise on information geometry (whose foundations depend on differential and Riemannian geometry). However, these geometric concepts are not well known in the economics literature, so we need to devote some considerable length for a brief intuitive overview of the key geometric ideas, and delegate the necessary technical details to the Appendix.

## 4.1 Information geometry

Consider a random variable $Z_p$ whose distribution is characterized by a parameterized density function [4] $f(z;p)$, where $p \in \Theta \subseteq \mathbb{R}^m$, and $\Theta$ is an open subset of $\mathbb{R}^m$. We assume that the function $p \mapsto f(\cdot;p)$ is smooth. We will also use the notation $f_p := f(\cdot;p)$ to emphasize the dependence of the distribution function on the parameterization rather than its function argument $z$. We collect all these distribution functions in a set $M := \{f_p : p \in \Theta\}$ and we call this set the *statistical model*.

The key observation in information geometry is to recognize that instead of directly working with $M$, which is a fairly complicated functional space, we work with its parameter space $\Theta$. However, for essentially all parameterized probability models, there is no such thing as an unique parameterization. That is to say, all models are identical up to reparameterization. If we identify all reparameterizations (e.g. $C^\infty$-diffeomorphisms) of the parameter space as identical to each other, then the statistical model $M$ can actually be viewed as a *statistical manifold*. The key ideas are illustrated in Figure 2.

We consider the Gaussian distribution to fix ideas and indeed, this will be the core model of our paper.

**Example 1** (Univariate Gaussian distribution)**.** Let $Z$ be a univariate Gaussian distribution with mean $\mu$ and standard deviation $\sigma$. Here, $\mathcal{Z} = \mathbb{R}$, $p = (\mu, \sigma)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{++}$. As it is

[4] We deliberately use the notation $p$ to denote the parameter of a distribution function, rather than with Greek letters like $\theta$. Our notation emphasizes the geometric nature of our framework and furthermore agrees with the conventions of differential geometry references.

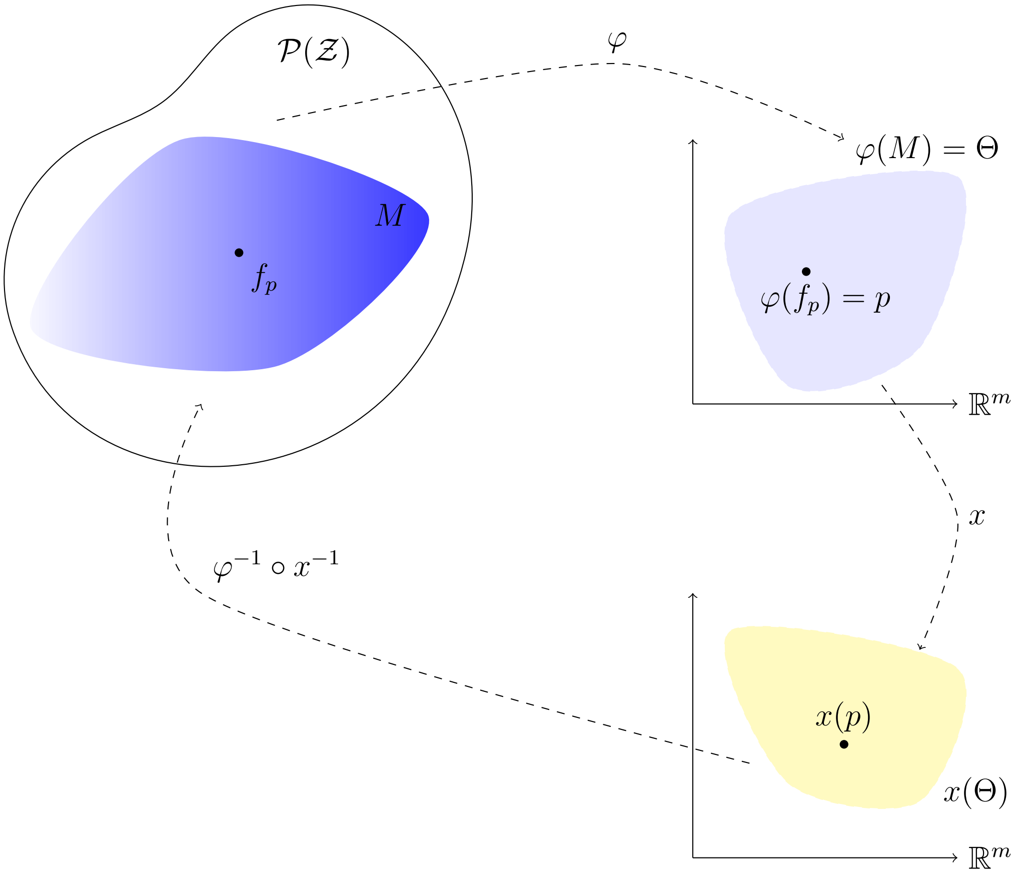

**Figure 2: Statistical manifold.** Let $\mathcal{P}(\mathcal{Z})$ define the space of probability density functions whose support is $\mathcal{Z}$ (e.g. we assume that the distribution support is parameter invariant). In particular, this means $\mathcal{P}(\mathcal{Z}) := \{f : \mathcal{Z} \to \mathbb{R} : f > 0 \,\text{and}\, \int_{\mathcal{Z}} f(z)dz = 1\}$. The space $\mathcal{P}(\mathcal{Z})$ is too big for our purposes, and hence we restrict to the subset $M$ of parameterized probability distributions. That is, let's consider probability distributions of the form $f(z; p)$ where $p \in \Theta \subseteq \mathbb{R}^m$, and so $M := \{f_p := f(\cdot; p) : p \in \Theta\}$ and this is the *statistical model*. Let's consider the mapping $\varphi : S \to \mathbb{R}^m$ given by $\varphi(f_p) = p$. Thus the mapping $\varphi$ identifies the probability distribution $f_p = f(z; p)$ with its parameterization $p$. Next, consider a $C^\infty$ diffeomorphism $x : \Theta \to \mathbb{R}^m$ and let $\rho = x(p)$. The map $x$ can thus be seen as a reparameterization of the distribution in the sense that the set $\{f_{\varphi^{-1} \circ x^{-1}(\rho)} : \rho \in x(\Theta)\}$ agrees with the statistical model $M$.

well known, its probability density function is

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

The *moment* parametrization $p = (\mu, \sigma)$ is fairly conventional in all elementary statistics references.

However, there is nothing inherently intrinsic to using the parametrization $p = (\mu, \sigma)$ to describe the Gaussian distribution. Indeed, consider the map $x : \Theta \to \mathbb{R}^2$ given by,

$$x(p) = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix} = \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}.$$

It is easy to verify that $x$ is a $C^\infty$-diffeomorphism. Using this parameterization, we can rewrite the density function as,

$$f(z;(x^1,x^2)) = \exp\left\{x^1 z + x^2 z^2 - \left(-\frac{(x^1)^2}{4x^2} + \frac{1}{2}\log(-\pi/x^2)\right)\right\}.$$

This is often called the *canonical* form of the Gaussian parameterization. Moreover, one can even analogously consider the map $y : \Theta \to \mathbb{R}^2$ given by,

$$y(p) = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix} = \begin{pmatrix} y^1 \\ y^2 \end{pmatrix},$$

and verify again that $y$ is a $C^\infty$-diffeomorphism. This is often called *expectation* form of the Gaussian parametrization.

Example 1 exemplifies that there is no “unique” way to describe the Gaussian distribution (or really any general distribution). Moreover, any $C^\infty$-diffeomorphism $x$ from one parametrization to another parameterization can equally describe the same Gaussian distribution. In our information acquisition context, this means that if we are to model an agent’s information mindset with Gaussian random variables, it must *not* be sensitive to how it is written. Information must intrinsically represent itself and not depend on how it is arbitrarily described or measured. The *manifold* formalism explicitly addresses the need to describe objects intrinsically.

## 4.2 Distance and geodesics on the statistical manifold

We introduce a notion of *distance* between two far points $p, q \in M$, and discuss the trajectory that connects them with minimal distance (*minimal geodesics*).

### 4.2.1 Distance and minimal cost of information acquisition

Recall our discussion of the relationship between Kullback-Leibler divergence and the Fisher information metric in Lemma 2.1. As discussed, the Fisher information metric approximates the KL divergence for two infinitesimally close random variables. Due to the nature of infinitesimal approximation, the Fisher information metric really only provides an *inner product* between the tangent vectors at a single point. Put simply, the Fisher information metric only measures the *angle* between two directions of information acquisition at a *given* point of knowledge $p \in M$. The Fisher information metric does not directly yield our desired distance between two distinct knowledge points $p, q \in M$. Again in our intrinsic theory of information acquisition, the measure of distance between two knowledge points proxies the cost of information acquisition from knowledge point $p$ to knowledge point $q$.

Riemannian geometry provides a canonical way to resolve this problem. Let $\gamma : [a, b] \to M$ be an arbitrary smooth trajectory in the statistical manifold $M$, meaning each $t \in [a, b] \subset \mathbb{R}$ is smoothly associated with a point $\gamma(t) \in M$. The *velocity* of the trajectory is defined as $\dot{\gamma}(t)$. As mentioned, the Fisher information metric $g$ is an inner product on the tangent space, and the velocity $\dot{\gamma}(t)$ is a tangent vector at the point $\gamma(t)$. This means we can define the *speed* of the trajectory as $||\dot{\gamma}(t)|| := \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}$. Since the *length* of a trajectory is the sum of its speed, we can define $L(\gamma) := \int_p^q ||\dot{\gamma}(t)||\, \mathrm{d}t$. In our context, $L(\gamma)$ represents the cost of information acquisition from an initial knowledge $p = \gamma(a)$ to a terminal knowledge $q = \gamma(b)$.

However, the agent is naturally interested in *minimizing* the cost of information acquisition. To this end, the *distance* between two points $p$ and $q$ can be defined to the length of the shortest trajectory between the two points. That is, the distance between $p, q$ can be defined as [5] ,

$$d(p, q) := \inf\{L(\gamma) : \gamma \text{ is a smooth curve in } M \text{ with } \gamma(a) = p \text{ and } \gamma(b) = q\ \}. \tag{4.1}$$

[5] It can be easily shown that in $\mathbb{R}^1$ with the Euclidean Riemannian metric $g_{ij} = \delta_{ij}$, the distance function between $x, y \in \mathbb{R}^1$ is $d(x, y) = |x - y|$, which is the usual Euclidean distance.

Thus, $d(p,q)$ is the *minimal* information acquisition cost from knowledge point $p$ to knowledge point $q$.

### 4.2.2 Raison d'être of our information geometric setup

At this point, one might wonder whether it is worth all the trouble to define this distance function $d$ to just measure the distance between two random variables? Recall that to arrive at (4.1), we had to approximate the Kullback-Leibler divergence to arrive at the Fisher information matrix, argue that the Fisher information matrix can be used as a metric, use this metric to measure angles between tangent vectors on the statistical manifold $M$, and then eventually push through several non-trivial arguments and definitions to define $d$. This entire process is admittedly complex. Why go through all this trouble to describe the distance between two knowledge points when the Kullback-Leibler divergence or even the mutual information as used by Sims (2003) provide a similar qualitative answer?

The true payoff of (4.1) to us is actually not the *value* $d(p,q)$, but rather its *trajectory solution*: what exactly is that trajectory $\gamma$ that solves the minimization problem $\inf_\gamma L(\gamma)$? This trajectory $\gamma$ tells us geometric qualities like direction, speed, acceleration on how an agent will acquire information from one point to another at minimal cost. These geometric qualities are simply not available if we use the Kullback-Leibler divergence or mutual information to just quantify the distance between two points of knowledge. As a result and to the best of my knowledge, existing papers in economics that use entropy-related measures are completely silent on the trajectories of information acquisition. These geometric qualities on the trajectory of information acquisition are precisely why we pursue an information geometric approach in this paper.

## 4.3 Geodesics: Solution to the minimal information acquisition cost problem

A key component of our paper is the trajectory solution $\gamma$ to the distance minimization problem (4.1). The trajectory solution to (4.1) actually plays a critical role in the development of Riemannian geometry, and hence they deserve the special name of *geodesics*. Roughly speaking, geodesics are "straight lines".

In a Euclidean space, a straight line can be described by a "zero acceleration" condition;

that is, if $\gamma(t)$ is a trajectory in $\mathbb{R}^n$, then $\gamma$ is a geodesic (i.e. a straight line) if satisfies the condition $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\gamma(t) = 0$. Furthermore, it is intuitive that the trajectory of the shortest distance between two points on a flat space are straight lines. Thus, geodesics *are* the solution to the minimum distance problem (4.1) in $\mathbb{R}^n$.

Analogously, the solution to (4.1) on our statistical manifold $M$ can be described by geodesics. However, trajectories that satisfy a "zero acceleration" condition in our manifold $M$ is far more nuanced due to the presence of non-zero *curvature* (we postpone an intuitive discussion of curvature to Section 4.4). Intuitively speaking, when a space is curved — as opposed to a flat space like the Euclidean space with zero curvature — the shortest distance between two points is *not* a "straight line" anymore. Geodesics on a general manifold $M$ cannot be described by the zero second derivative condition $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\gamma(t) = 0$. Unlike a flat Euclidean space, non-zero curvature causes tangent spaces at different points on the manifold to be incomparable. Instead of flat space differentiation $\frac{\mathrm{d}}{\mathrm{d}t}$, we must develop the idea of *covariant differentiation* $\nabla$ to describe the zero acceleration condition in a curved manifold.

In all, the *minimal geodesic* $\gamma$ that solves (4.1) is actually the solution to the *geodesic equation* $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, and here $\dot{\gamma}(t)$ a tangent vector at the point $\gamma(t) \in M$. We have a more technical overview of geodesics in Section A.3. Before we present the concrete solution to (4.1), we will discuss the role of curvature in our context in Section 4.4.

## 4.4 Intuitive discussion of curvature

The heart of Riemannian geometry is *curvature* [6] . Indeed, essentially *all* of the results in our paper are driven by the curvature of the agent's statistical manifold. Unfortunately, it is far beyond the scope of this paper to present a full technical discussion (see Lee (1997) and Petersen (2016) for an introductory treatment). Here we will present two different perspectives on curvature. In Section 4.4.1, we show that if relative entropy measures the information processing cost between two different pieces of knowledge, then the curvature in the agent's statistical manifold arises because the cost is dependent on the agent's current knowledge position. As a result, the agent might find it more or less costly to turn to acquire information in one knowledge direction versus another. It is this direction "turning" of

[6] Indeed, the heart of Einstein's general theory of relativity is the curvature of spacetime. Einstein spent eight years (from 1907 to 1915) to develop his theory. He struggled many of those years in connecting the concept spacetime curvature to observable physical phenomenon.

information acquisition generates curvature. In Section 4.4.2 we discuss how our information geometric framework relates to a noisy signal extraction framework that's used in many applied economic theory models. In Section 4.4.3, we introduce the concept of *information capacity* and present a more geometric perspective on curvature. Information capacity is the amount of similar knowledge around a given point of knowledge. Information capacity is high (low) when an econometrician can (cannot) easily infer the knowledge of an informed agent. We show that high (low) curvature leads to low (high) information capacity. This perspective on curvature thus lends to potential empirical implementations of our framework.

Finally, we note that for ease of exposition, all of the discussions of curvature in this section are explicitly coordinate dependent. Recall our principles for an intrinsic theory of information acquisition have emphasized that knowledge should not be dependent on the coordinate system. We make clear that all of the discussions in this section can be made precise in a coordinate invariant fashion using tensor calculus.

### 4.4.1 Knowledge dependent marginal cost

Mathematically, the curvature of a Riemannian manifold is entirely driven by its Riemannian metric $g$, and in our case, the Fisher information metric. Economically, however, it might seem a bit mysterious as to why curvature so prominently drives most of our information acquisition results. In a nutshell, curvature arises because the speed and direction of acquiring knowledge in one direction depends on the position of another piece of knowledge.

To give a better intuitive understanding of the source of curvature, let's revisit the Kullback-Leibler divergence and recall that in Lemma 2.1, the Fisher information metric is the infinitesimal version of the KL divergence. Let $\mathcal{N}_0 := \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}_1 := \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ be two $m$-dimensional Gaussian distributions. Then the KL-divergence between them is,

$$D_{KL}(\mathcal{N}_1 || \mathcal{N}_0) = \frac{1}{2}\left(\operatorname{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + \log\left(\frac{\det \boldsymbol{\Sigma}_0}{\det \boldsymbol{\Sigma}_1}\right) - m\right). \quad (4.2)$$

As discussed in Section 2.1 and 2.2, this relative entropy between two states of knowledge represents the distance of two states of knowledge for our agent.

For the sake of discussion, let's intuitively regard "naive" comparative statics on (4.2) as "change in knowledge" from knowledge state $\mathcal{N}_0$ to $\mathcal{N}(\boldsymbol{\mu}_1 + \mathrm{d}\boldsymbol{\mu}, \boldsymbol{\Sigma}_1 + \mathrm{d}\boldsymbol{\Sigma})$. That is, let's regard $\frac{\partial}{\partial(\boldsymbol{\mu}_1)^k}$ as the change in knowledge in the $k$th component in the $\boldsymbol{\mu}_1$ direction, and regard

$\frac{\partial}{\partial(\mathbf{\Sigma}_1)^{ij}}$ as the change in knowledge in the $(i,j)$th component in the $\mathbf{\Sigma}_1$ direction. Holding knowledge state $\mathcal{N}_0$ as fixed, it is clear that both the qualitative and quantitative effects of $\frac{\partial}{\partial(\boldsymbol{\mu}_1)^k}$ and $\frac{\partial}{\partial(\mathbf{\Sigma}_1)^{ij}}$ are generically different, and heavily depend on the current knowledge state $\mathcal{N}_0$ value. This is to say, the effect of changing knowledge in the direction of the mean is substantially different from changing knowledge in the direction of the variance-covariance.

To make this point even clearer, let's consider $m = 1$ dimensional Gaussian distributions with $\mathcal{N} = \mathcal{N}(\mu, v)$ (mean $\mu$, variance $v$) and $\mathcal{N}_1 = \mathcal{N}(\mu_1, v_1)$, and $\mu_1 = \mu + \mathrm{d}\mu, v_1 = v + \mathrm{d}v$. Their KL divergence is

$$\begin{aligned} D_{KL}(\mathcal{N}_1 || \mathcal{N}) &= \frac{1}{2}\left(\frac{v_1}{v} + \frac{(\mu-\mu_1)^2}{v} + \log\frac{v}{v_1} - 1\right) \\ &\approx \frac{1}{v}(\mu-\mu_0)^2 + \frac{1}{2v^2}(v_1 - v)^2 \\ &= \frac{1}{v}\mathrm{d}\mu^2 + \frac{1}{2v^2}\mathrm{d}v^2 \\ &= \mathrm{d}s^2, \end{aligned} \tag{4.3}$$

where the approximation is a second order Taylor expansion of $(\mu_1, v_1)$ around $(\mu, v)$. The symbol $\mathrm{d}s$ is the standard notation for the infinitesimal length of a line element, and so $\mathrm{d}s^2$ is the squared line element. This notation enforces the geometric nature of our framework, that measuring the distance between knowledge is akin to measuring the distance between physical locations [7] .

We are ready to intuitively understand why curvature arises in our information geometric framework. Thinking of the KL divergence or relative entropy as the cost of limited capacity of information processing for an agent, we see from (4.3) that if the current knowledge is $(\mu, v)$, then acquiring knowledge in the mean-only direction incurs a marginal cost of $1/v$, while acquiring knowledge in the variance-only direction incurs a marginal cost of $1/(2v^2)$.

[7] It is useful to think about this notation in $\mathbb{R}^3$. In Cartesian coordinates, the distance between a point $(x, y, z) \in \mathbb{R}^3$ and $(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z) \in \mathbb{R}^3$ is $\mathrm{d}s = \sqrt{(x-(x+\mathrm{d}x))^2 + (y-(y+\mathrm{d}y))^2 + (z-(z+\mathrm{d}z))^2} = \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2 + (\mathrm{d}z)^2}$. As a matter of notation, it is customary to write $\mathrm{d}x\mathrm{d}x$ as $\mathrm{d}x^2$, and not $(\mathrm{d}x)^2$. Using this notation and rearranging, the squared line element in $\mathbb{R}^3$ under Cartesian coordinates is $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. Here we see that there are no leading coefficients in the terms $\mathrm{d}x^2, \mathrm{d}y^2, \mathrm{d}z^2$. This is in contrast to (4.3) where there are explicit leading coefficients present in front of $\mathrm{d}\mu^2$ and $\mathrm{d}v^2$. This is not an unique phenomenon to our setup. For instance, the same line element in $\mathbb{R}^3$ in spherical polar coordinates $(r, \theta, \phi)$ (radius, polar angle and azimuth) is $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2$, and the leading coefficients are explicitly here.

Observe that $1/v$ is larger (smaller) than $1/(2v^2)$ for large (small) values of $v$. This implies the marginal cost of acquiring knowledge in a particular direction is heavily dependent on the position of the current knowledge. This is what generates *curvature* in information acquisition trajectories — curvature is present because the agent will find it less costly to acquire knowledge quicker or slower in one direction versus another. Moving in one direction quicker or slower necessarily requires the agent to "turn directions", resulting in curved trajectories.

### 4.4.2 Relationship to noisy signal information acquisition frameworks

How does our framework relate to the conventional signal plus noise information acquisition framework? See for instance Grossman and Stiglitz (1980) and many others for examples of this framework. In that framework, an agent observes a signal $U$ that is given by $U = S + \varepsilon$, where $S \sim \mathcal{N}(0, 1)$ is the unobservable variable of interest, and $\varepsilon \sim \mathcal{N}(0, \hat{v})$ is a noise term with *known* variance $\hat{v}$. Assume that $S$ and $\varepsilon$ are independent. The agent acquires information $S$ through the observation of signal $U$. How does the variance $\hat{v}$ of the noise term affect the agent's signal $S$? Intuitively, the variable of interest $S$ becomes more clearly revealed to the agent when the variance $\hat{v}$ of the noise is smaller, or when the precision $1/\hat{v}$ is higher. Let's make this statement precise and see how it relates to our information geometric framework. Since $S$ and $\varepsilon$ are Gaussian, it is straightforward to show that upon seeing the signal $U = u$, the agent infers that the variable of interest is $\mathbb{E}[S|U = u] = \frac{1}{1+\hat{v}}u$. Suppose the agent compares the signal $U = u$ versus a small variation of it $U = u + \mathrm{d}u$. Then the difference in inferring $S$ between the two signal observations is $\mathrm{d}s = \mathbb{E}[S|U = u + \mathrm{d}u] - \mathbb{E}[S|U = u] = \frac{1}{1+\hat{v}}\mathrm{d}u$. This term $\mathrm{d}s = \frac{1}{1+\hat{v}}\mathrm{d}u$ can be interpreted as the "distance" between two nearby signal means, and we see that this distance is proportional to the *known* precision $1/(1 + \hat{v})$ of the noise.

Our information geometric framework has an analogous counterpart to the aforementioned noisy signal framework. Suppose the variance $\hat{v}$ of the truth information $\hat{Z}$ is known to the agent at $t = 0$. The agent only needs to acquire information of the mean $\mu$. In this case, the single element Fisher information matrix is $g_{ij}(p) = [1/\hat{v}]$. This implies the line element takes the form

$$\mathrm{d}s^2 = \frac{1}{\hat{v}}\mathrm{d}\mu^2, \tag{4.4}$$

or that $\mathrm{d}s = \frac{1}{\sqrt{\hat{v}}}\mathrm{d}\mu$. The distance between two pieces of mean information is proportional to the *known* precision of the information. We see that up to a monotonic transformation on the precision term, the distance between information in the above noisy signal framework has a direct analogy to our information geometric framework.

The strength of an information geometric setup is that we can analyze information acquisition when different types of knowledge are known or unknown. In the above two discussions for the distance between information, a key assumption is that the precision is known. However, it is not straightforward to alter the noisy signal framework when the precision is also unknown. In contrast, our information geometric framework is well suited to handle all three possible cases in discussing the distance between information: (i) both the mean and variance are unknown; (ii) the mean is unknown and the variance is known; and (iii) the mean is known and the variance is unknown.

We had already shown case (i) in (4.3), and case (ii) in (4.4). For case (iii), suppose the mean $\hat{\mu}$ of the truth information $\hat{Z}$ is known to the agent at $t = 0$. The agent only needs to acquire information of the variance $v$. In this case, the single element Fisher information matrix is $g_{ij}(p) = [1/(2v^2)]$. As a result, the line element is,

$$\mathrm{d}s^2 = \frac{1}{2v^2}\mathrm{d}v^2. \tag{4.5}$$

Interestingly, the known mean value $\hat{\mu}$ does not enter into the distance between knowledge of the variance. In contrast, in the case (4.4) when the mean is unknown but the variance is known, the distance between knowledge of the mean explicitly depends on the known variance.

### 4.4.3 Information capacity

We present also a geometric argument to intuitively understand curvature. We will argue a notion of *information capacity* in our economy is completely determined by the *(scalar) curvature* of the statistical manifold.

In probability theory, the *Jeffreys prior* $f^J$ is a non-informative (objective) prior distribution of the statistical manifold at points $p \in M$. It is defined to be $f^J(p) \propto \sqrt{\det[g_{ij}(p)]}$, where the right hand side is the square root of the determinant of the Fisher's information matrix at $p$. In our context, it is useful to think of the Jeffreys prior to represent the knowledge states of an *uninformed outsider* who only knows an objective *distribution* of the points $p \in M$, but not their exact location. It may also be convenient to think of this uninformed outsider as an econometrician who has to statistically estimate and infer the knowledge $p \in M$ of an agent based on observable data.

In Riemannian geometry and in coordinates, one defines the *volume form*

$$\omega := \sqrt{\det[g_{ij}(p)]}\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m \tag{4.6}$$

of an orientable manifold $M$, and where $p = (x^1(p), \ldots, x^m(p))$. Thus, the Jeffreys prior is directly proportional to the volume form $f^J(p)\mathrm{d}x^1 \cdots \mathrm{d}x^m \propto \omega$. Let $J$ be a random variable with the probability density function $f^J$, and consider a small open (geodesic) ball $B_M(p,r)$ at $p$ with radius $r$. The probability $\mathbb{P}(J \in B_M(p,r))$ can be regarded the likelihood of an uninformed outsider "guessing" correctly an informed agent's knowledge at $p \in M$ within a small window of error $r$. We make the following observation,

$$\mathbb{P}(J \in B_M(p,r)) = \int_{B_M(p,r)} f_J(p)\,\mathrm{d}x^1 \cdots \mathrm{d}x^m \propto \int_{B_M(p,r)} \omega = \mathrm{vol}(B_M(p,r)). \tag{4.7}$$

From (4.7) and our identification of points $p \in M$ as states of knowledge, the volume $\mathrm{vol}(B_M(p,r))$ can be labelled as the *information capacity* at a point $p$ with radius $r$. Information capacity is high when an uninformed outsider has a high likelihood of guessing the knowledge $p$ with some error $r$. Conversely, information capacity is low when an uninformed outsider has a difficult time of guessing the knowledge $p$ possessed by an informed agent.

Now, we relate the information capacity to *curvature*. Gray (1974) shows,

$$\mathrm{vol}(B_M(p,r)) \approx \mathrm{vol}(B_{\mathbb{R}^m}(r)) \left(1 - \frac{S}{6(m+2)}r^2\right), \tag{4.8}$$

where $S$ is the *scalar curvature* of $M$ at point $p$, and $B_{\mathbb{R}^m}(r)$ is a Euclidean ball [8] of radius $r$ in $\mathbb{R}^m$. Equation 4.8 shows the volume of $B_M(p,r)$ is proportional to an $m$-dimensional volume ball $B_{\mathbb{R}^m}(r)$ of the same radius in Euclidean space, but scaled by the factor $1 - \frac{S}{6(m+2)}r^2$. Naturally, the volume of $B_M(p,r)$ is dependent on the dimension $m$ of the manifold $M$. But

[8] The volume of the Euclidean ball is invariant to its location, and indeed, $\mathrm{vol}(B_{\mathbb{R}^m}(r)) = \alpha_m r^m/m$, where $\alpha_m := 2\frac{\Gamma(1/2)^m}{\Gamma(m/2)}$, and $\Gamma(z) := \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t$ is the gamma function.

importantly, $S$ is the *scalar curvature* of manifold $M$ at $p$, and it is a real valued scalar. Hence, *positive* scalar curvature $S > 0$ *decreases* volume, *zero* scalar curvature $S = 0$ means the volume in $M$ is identical to that in $\mathbb{R}^m$, and *negative* scalar curvature $S < 0$ *increases* volume. In all, *curvature* is the concrete determinant of information capacity in our economy. [9]

# 5 Gaussian information geometry

Having presented the intuitive background of differential geometry, Riemannian geometry and information geometry, we now present the key application of these concepts in this paper — the information geometry of the Gaussian distribution.

Let $M$ be the manifold of associate with an $n$-dimensional Gaussian distribution. Since an $n$-dimensional Gaussian distribution can be characterized by $m = n + n(n+1)/2$ number of parameters, this means $n$-dimensional Gaussian distributions are associated with an $m$-dimensional manifold $M$. Furthermore, pursuant to the motivations in Section A we equip $M$ with the Fisher information metric [10] as the Riemannian metric $g$.

At this point, we need to address an important issue of existence and uniqueness of the geodesics $\gamma$ that solves (4.1). These are actually deep and difficult questions in Riemannian geometry. However, we can give the following result for the case of Gaussian statistical manifolds, which will be the focus of application in our paper.

**Proposition 5.1** (Minimizing geodesics in Gaussian statistical manifold)**.** *Let $M$ represent a Gaussian statistical manifold. Then there always exists a* minimizing *geodesic connecting two points $p, q \in M$.*

[9] Some fundamental results in Riemannian geometry actually say a lot more about manifolds of *constant* curvature. In our setup here, $S$ need not be constant for all points $p \in M$. But if $S$ is identically constant on the manifold, then fundamental results in Riemannian geometry show that: when $S = 0$, the manifold is isomorphic to the flat Euclidean space; when $S > 0$, the manifold is isomorphic to a high dimensional sphere; and when $S < 0$, the manifold is isomorphic to a hyperboloid.

[10] As remarked earlier, this paper shares a lot of mathematical techniques used in general relativity. However, there is one important difference and it boils down to how one comes about the Riemannian metric. For instance in general relativity, the metric on spacetime is a *result* from the famous Einstein's field equations, and these equations embed all the physical laws of interest. In our information geometric context, the metric is *given* to be the Fisher information metric, of which the microeconomic foundations of this choice have been discussed in Section A.

*Proof of Proposition 5.1.* The full argument is a little bit nuanced (see Calvo and Oller (1990) for details) but here we'll sketch the main ideas. Firstly, there is a single global atlas associated with an $m$-dimensional Gaussian statistical manifold $M$. In particular, let's pick that global atlas to have mean-variance-covariance coordinates $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $M$ can be embedded diffeormophically and isometrically into a manifold $P$ of positive definite matrices with the Siegel Riemannian metric. The manifold $P$ with the Siegel metric is a *geodesically complete* manifold. By the Hopf-Rinow theorem (see Lee (1997)), this means any two points in $P$ can be joined with minimal distance (with respect to the Siegel metric). Finally, since $P$ and $M$ are diffeomorphic and isometric to each other, this means any two points in $M$ can be joined with minimal distance (with respect to $d$ on $M$). □

The important implication of Proposition 5.1 is that there *always* exist a trajectory of minimal cost to acquire information from one point of knowledge to another. Notice however the result does not claim uniqueness [11] .

Let's begin with our discussion from an univariate Gaussian distribution and then generalize to multivariate Gaussian distributions. As we shall see, the dimensionality makes critical qualitative and quantitative differences to how an agent acquires information. In this section, we will interchange between mean-precision coordinates and mean-variance-covariance coordinates in describing the Gaussian distribution, depending on which coordinate system is more convenient for exposition purposes. In the economic applications of Section 6 and Section 7, we will primarily use mean-precision coordinates for computational ease.

## 5.1 Univariate Gaussian distribution

We summarize the main geometric results for the statistical manifold of an univariate Gaussian distribution.

**Proposition 5.2** (Geometric properties of an univariate Gaussian statistical manifold)**.** *Let an $m = 2$ dimensional manifold $M$ represent the statistical manifold of an univariate Gaussian distribution. Then in mean-precision coordinates, $\varphi(p) = (x^1, x^2) = (\mu, \lambda)$,*

[11] As an illustration to see why uniqueness is generally not attainable, suppose one wants to travel from the North pole of the earth to the South pole. Any great circle (aka. *orthodome*) connecting between the North pole and the South pole will be of minimal distance. But there are infinitely many great circles between the two poles.

*(i) The Fisher information metric $g$ at point $p \in M$ is,*

$$[g_{ij}(p)] = \begin{pmatrix} \lambda & \\ & \frac{1}{2\lambda^2} \end{pmatrix} = \begin{pmatrix} x^2 & \\ & \frac{1}{2(x^2)^2} \end{pmatrix}, \tag{5.1}$$

*where the blank entries are zeros.*

*(ii) The geodesic equation on $M$ is,*

$$\ddot{\mu} = -2\Gamma^1_{21}\dot{\lambda}\dot{\mu} = -\frac{\dot{\lambda}\dot{\mu}}{\lambda}, \tag{5.2a}$$

$$\ddot{\lambda} = -\Gamma^2_{11}\dot{\mu}^2 - \Gamma^2_{22}\dot{\lambda}^2 = \frac{\lambda^3\dot{\mu}^2 + \dot{\lambda}^2}{\lambda}. \tag{5.2b}$$

*The $\Gamma^k_{ij}$ are the* Christoffel symbols*, for $i, j, k = 1, \ldots, m$, and they are given by,*

$$\Gamma^1_{21} = \frac{1}{2\lambda}, \tag{5.3a}$$

$$\Gamma^2_{11} = -\lambda^2, \tag{5.3b}$$

$$\Gamma^2_{22} = -\frac{1}{\lambda}, \tag{5.3c}$$

*and where all other non-displayed $\Gamma^k_{ij}$'s are equal to zero.*

*(iii) The explicit solution to the geodesics* (5.2) *is either (a):*

$$\mu(t) = c_1,$$
$$\lambda(t) = \exp\{-\sqrt{2}t + c_0\},$$

*or (b):*

$$\mu(t) = c_1 + 2c_2 \tanh(t/\sqrt{2} + c_0),$$
$$\lambda(t) = \frac{\cosh^2(t/\sqrt{2} + c_0)}{2c_2^2}$$

*where $c_0, c_1, c_2$ are real constants.*

*(iv) For two points $p = (\mu_p, \lambda_p)$ and $q = (\mu_q, \lambda_q)$, the geodesic distance* (4.1) *is,*

$$d(p,q) = d_{\mathbb{H}}\left(\left(\frac{\mu_p}{\sqrt{2}}, \frac{1}{\sqrt{\lambda_p}}\right), \left(\frac{\mu_q}{\sqrt{2}}, \frac{1}{\sqrt{\lambda_q}}\right)\right), \tag{5.6}$$

*where $d_{\mathbb{H}}$ is the geodesic distance is on the Poincaŕe half-plane,*

$$d_{\mathbb{H}}((x^1, y^1), (x^2, x^2)) = \operatorname{arcosh}\left(1 + \frac{(x^2 - x^1)^2 + (y^2 - y^1)^2}{2y^1y^2}\right).$$

*Proof of Proposition 5.2.* The proof arguments in (i) and (ii) are general. The proof arguments in (iii) and (iv) are special to the univariate Gaussian distribution.

Part (i) follows immediately from the definition of the Fisher information matrix,

$$g_{ij}(p) := \mathbb{E}_p\left[\left(\partial_i \log f(Z;x)|_p\right)\left(\partial_j \log f(Z;x)|_p\right)\right], \tag{5.7}$$

where in this case $Z$ is an univariate Gaussian distribution with mean $\mu$ and precision $\lambda$, and $f$ is the associated probability density function.

Part (ii) is an immediate application of the general formula (A.2) for the geodesic equation in local coordinates.

Let's see part (iii). We note that Calvo and Oller (1991) show the explicit solution for the geodesics of multivariate Gaussians. The solution presented here is directly from Skovgaard (1984).

Part (iv) is quite special to the univariate Gaussian distribution (or slightly more generally, multivariate Gaussian distribution with different mean elements, but identical variances and zero correlation). Rather than solving the optimization problem (4.1) directly, we appeal to a geometric argument. A direct calculation (see Amari and Nagaoka (2007) and Amari (2016)) shows that the statistical manifold $M$ for the univariate Gaussian distribution has a scalar curvature of $-1$. From standard Riemannian geometry (see Lee (1997)), manifolds with a scalar curvature of $-1$ can be identified with the hyperbolic geometry. In particular, the Poincaŕe half-plane $\mathbb{H}^2 := \{(x,y) : x \in \mathbb{R}, y \in \mathbb{R}_{++}\}$ is the canonical two-dimensional manifold with hyperbolic geometry. In all, this means up to coordinate transformations, the geodesic distance properties of $M$ is similar to that of $\mathbb{H}^2$.

The Riemannian metric on $\mathbb{H}^2$ at a point $p = (x, y)$,

$$[g_{ij}(p)]_{\mathbb{H}} = \begin{bmatrix} \frac{1}{y^2} & \\ & \frac{1}{y^2} \end{bmatrix}.$$

From properties of $\mathbb{H}^2$, the geodesic distance $d_{\mathbb{H}}$ between two points $(x^1, y^1)$ and $(x^2, y^2)$ is

$$\begin{aligned} d_{\mathbb{H}}((x^1, y^1), (x^2, x^2)) &= \operatorname{arcosh}\left(1 + \frac{(x^2 - x^1)^2 + (y^2 - y^1)^2}{2y^1y^2}\right) \\ &= 2\log\frac{\sqrt{(x^2 - x^1)^2 + (y^2 - y^1)^2} + \sqrt{(x^2 - x^1)^2 + (y^2 + y^1)^2}}{2\sqrt{y^1y^2}} \end{aligned}$$

Compare the Fisher information matrix $g$ in mean-standard deviation coordinates as shown above, and the Riemannian metric on $\mathbb{H}^2$. We see that by considering a similarity mapping $F : M^2 \mapsto \mathbb{H}^2$ defined by $F(\mu, \sigma) = (\mu/\sqrt{2}, \sigma)$, we obtain that,

$$d((\mu_1, \sigma_1), (\mu_2, \sigma_2)) = d_{\mathbb{H}}\left((\mu_1/\sqrt{2}, \sigma_1), (\mu_2/\sqrt{2}, \sigma_2)\right).$$

However, we are most interested in geodesic distance expressed in mean-precision variables. And since $(\mu, \sigma) \mapsto (\mu, 1/\sqrt{\lambda})$, the geodesic distance in mean-precision variables is given by,

$$d((\mu_1, 1/\sqrt{\lambda_1}), (\mu_2, 1/\sqrt{\lambda_2})) = d_{\mathbb{H}}\left((\mu_1/\sqrt{2}, 1/\sqrt{\lambda_1}), (\mu_2/\sqrt{2}, 1/\sqrt{\lambda_2})\right).$$

□

**Corollary 5.3** (Geodesic connecting two points on univariate Gaussian manifold)**.** *Suppose we have two points, starting at $t = 0$ with $p = (\mu_p, \lambda_p)$ and ending at $t = T$ with $q = (\mu_q, \lambda_q)$ expressed in mean-precision coordinates on an univariate Gaussian statistical manifold.*

*(i) If $\mu_p = \mu_q$, then the minimal geodesic that connects $p$ to $q$ follow case (a) in Proposition 5.2(iii) with constants,*

$$\begin{aligned} c_0 &= \log\lambda_p \\ c_1 &= \mu_p \\ T &= \frac{1}{\sqrt{2}}\log\frac{\lambda_p}{\lambda_q}. \end{aligned}$$

*(ii) If $\mu_p \neq \mu_q$, then the minimal geodesic that connects $p$ to $q$ follow case (b) in Proposition 5.2(iii) with constants,*

$$\begin{aligned}
c_0 &= \operatorname{arcosh}\left(\sqrt{2\lambda_p c_2^2}\right) = c_0(c_2) \\
c_1 &= \mu_p - 2c_2 \tanh(c_0) \\
T &= \sqrt{2}\left[\operatorname{arctanh}\left(\frac{2c_2\tanh(c_0) + \mu_q - \mu_p}{2c_2} - c_0\right)\right],
\end{aligned}$$

*and $c_2$ is the real solution to,*

$$0 = \cosh^2\left(\operatorname{arctanh}\left(\frac{2c_2\tanh(c_0) + \mu_q - \mu_p}{2c_2}\right)\right) - 2c_2^2\lambda_q.$$

*Proof of Corollary 5.3.* Let's begin with part (i). Clearly, since $\mu_p = \mu_q$, then we must have that $\mu(t) = c_1 = \mu_p$ for all $t$. At $t = 0$, we have that $\lambda_p = \lambda(0) = e^c_0$, and solving yield $c_0$. At $t = T$, we have that $\lambda_q = \lambda(T) = e^{-\sqrt{2}T + \log\lambda_p}$. Solving, we get $T$ as desired.

Let's consider part (ii). Evaluating at $t = 0$ and $t = T$, we have the following four equations to the boundary value problem,

$$\mu_p = \mu(0) = c_1 + 2c_2\tanh(c_0) \tag{5.8a}$$

$$\lambda_p = \lambda(0) = \frac{\cosh^2(c_0)}{2c_2^2} \tag{5.8b}$$

$$\mu_q = \mu(T) = c_1 + 2c_2\tanh(T/\sqrt{2} + c_0) \tag{5.8c}$$

$$\lambda_q = \lambda(T) = \frac{\cosh^2(T/\sqrt{2} + c_0)}{2c_2^2} \tag{5.8d}$$

Immediately, from (5.8b) we get $c_0$, and from (5.8a) we get $c_1$ as displayed. Substituting $c_1$ into (5.8c), we get that,

$$T/\sqrt{2} + c_0 = \operatorname{arctanh}\left(\frac{2c_2\tanh(c_0) + \mu_q - \mu_p}{2c_2}\right), \tag{5.9}$$

and we solve for $T$ and get the displayed expression. Substituting (5.9) into (5.8d), we get the displayed expression for the equation for $c_2$.

□

From the Fisher information metric, and using the line element notation, we can write $\mathrm{d}s^2 = \lambda \mathrm{d}\mu^2 + \frac{1}{2\lambda^2}\mathrm{d}\lambda^2 = x^2\mathrm{d}(x^1)^2 + \frac{1}{2(x^2)^2}\mathrm{d}(x^2)^2$. Notice that the line element in mean-precision coordinates differs from (4.3) in mean-variance coordinates. All geometric information about the manifold $M$ is encapsulated in the Fisher information metric $g$. Indeed, the geodesic equation (5.2) and the geodesic distance (5.6) can all be derived from knowing $g$ alone.

Recall from Section 4.2, the geodesics on a statistical manifold $M$ represent the optimal information acquisition trajectory at minimal cost. The geodesic equations (5.2) are precisely the trajectories that connect some initial knowledge point $p \in M$ to a terminal knowledge point $q \in M$. See Figure 3 for an illustration.

As discussed in Section 4.4, curvature is the key geometric driver of our framework. The *Christoffel symbols* $\Gamma^k_{ij}$'s of (5.3) have numerous important applications, but for our purposes, these symbols allow us to explicitly quantity and concretely introduce curvature. Observe that if the $\Gamma^k_{ij}$'s were equal to zero, then we have the geodesics $\ddot{\mu} = 0$ and $\ddot{\lambda} = 0$. These of course correspond to the Euclidean straight lines $\mu(t) = a_0 + a_1 t$ and $\lambda(t) = b_0 + b_1 t$.

As discussed in Section A.3, we can view geodesics both as straight lines or as the shortest trajectory between two points. In the straight lines perspective, a unique description of the geodesics is complete once we specify the initial position and initial velocity. In our information context, this means once we fix the agent's initial knowledge level $(\mu(0), v(0))$, and fix the agent's initial direction of information acquisition $(\dot{\mu}(0), \dot{v}(0))$, the agent's future $t > 0$ information acquisition trajectory is determined. The initial information acquisition direction $(\dot{\mu}(0), \dot{v}(0))$ can be thought of as the *optimism* or *pessimism* of information. For instance, if $\dot{\mu}(0) > 0$ then the agent is optimistic about the mean of the information relative to his initial knowledge, whereas if $\dot{\mu}(0) < 0$ then the agent is pessimistic. In this perspective, we work with an *initial value problem* for the geodesics.

The alternative perspective that geodesics are the shortest trajectories between two points will allow us to describe how the agent should learn the truth. This is the core motivation of Lemma 2.3 and indeed, this is its proof. Suppose the agent is endowed with some initial knowledge $Z(0)$ that we represent on the manifold $M$ as $p = (\mu(0), v(0))$. News arrive slowly to the agent. The agent must choose an information acquisition trajectory $Z(t)$ to arrive at the truth $\hat{Z}$ at some terminal time $T$. Let's represent the truth knowledge by the point $q = (\hat{\mu}, \hat{v})$ on the manifold. Then the trajectory $\{Z(t)\}_{t\in[0,T]}$ satisfying (5.2) that starts at $p$

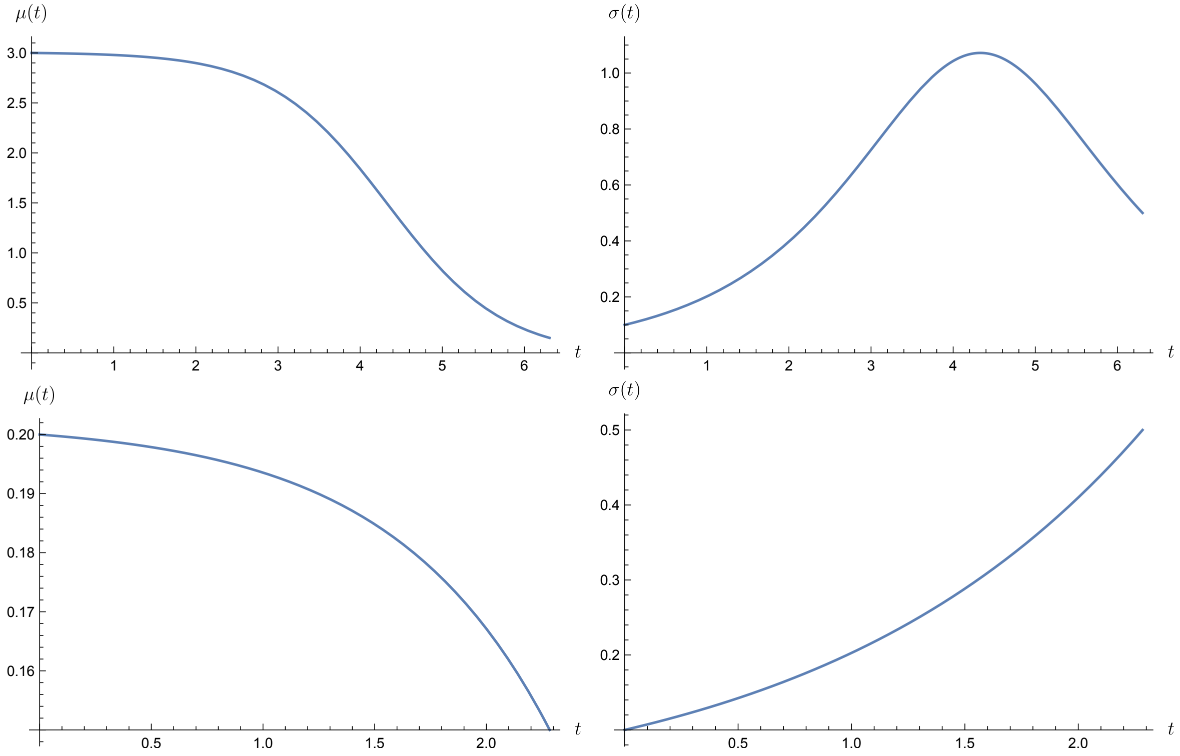

**Figure 3: Illustrations of the geodesics of an univariate Gaussian statistical manifold.** We numerically illustrate the equation solutions (5.2). For ease of presentation, we define the trajectory $\sigma(t) := 1/\sqrt{\lambda(t)}$. The top two figures are illustrated with parameter set (1): $\mu_0 = 3, \mu_1 = 0.15, \sigma_0 = 0.10, \sigma_1 = 0.5$. The bottom two figures are illustrated with parameter set(2), and they are the same set parameters as (1), except we replace the value $\mu_0 = 0.20$. The behavior of the geodesics can be quite different by just changing the initial value.

and ends at $q$ is the optimal information acquisition trajectory. In this perspective, we work with a *boundary value problem* for the geodesics.

## 5.2 Multivariate Gaussian distribution

The hard work for setting up the geometric structure of our framework truly pays off when one compares the geodesics of a multivariate versus a univariate Gaussian distribution. In vast majority of information economic models, the dimensionality of the signal in a Gaussian context usually is not a deciding factor in driving the qualitative economic results. However in our context, dimensionality matters significantly both qualitatively and quantitatively.

Again, let's work with the mean and precision [12] coordinates for a point $p$ as $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = (x^1, \ldots, x^m)$. Then it is easy to show that,

$$\begin{aligned}
\mathbb{E}_p\left[\frac{\partial^2 \log f(Z;p)}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\top}\right] &= \boldsymbol{\Lambda}, \\
\mathbb{E}_p\left[\frac{\partial^2 \log f(Z;p)}{\partial\boldsymbol{\mu}\partial\boldsymbol{\Lambda}}\right] &= \mathbf{0}, \\
\mathbb{E}_p\left[\frac{\partial^2 \log f(Z;p)}{\partial\boldsymbol{\Lambda}\partial\boldsymbol{\Lambda}}\right] &= \frac{1}{2}\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1},
\end{aligned}$$

and here $\otimes$ denotes the Kronecker product.

To be concrete, let's focus on a bivariate Gaussian distribution. Then a point $p$ can be written in local coordinates as $(\mu_1, \mu_2, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}) = (x^1, x^2, x^3, x^4, x^5, x^6)$. The Fisher information matrix is [13]

[12] Precision in this multivariate context is defined to be the inverse of the covariance-variance matrix.

[13] As we can see even in the bivariate case, parameterizing the Gaussian distribution in terms of precision facilitates for a (relatively) clean expression of the Fisher information matrix. If we were to parameterize using the covariance-variance matrix, the associated Fisher information matrix would be far more messy. However, as we have emphasized throughout the paper, the Riemannian metric is tensorial and so its geometric properties are invariant to coordinate choices. But in terms of practical computations, it may easier using one particular coordinate system over another.

$$[g_{ij}(p)] = \begin{pmatrix} \lambda_{11} & \lambda_{12} & & & & \\ \lambda_{21} & \lambda_{22} & & & & \\ & & \frac{\lambda_{22}^2}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{12}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{12}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{12}^2}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} \\ & & -\frac{\lambda_{21}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{11}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{12}\lambda_{21}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{11}\lambda_{12}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} \\ & & -\frac{\lambda_{21}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{12}\lambda_{21}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{11}\lambda_{22}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{11}\lambda_{12}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} \\ & & \frac{\lambda_{21}^2}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{11}\lambda_{21}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & -\frac{\lambda_{11}\lambda_{21}}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} & \frac{\lambda_{11}^2}{(\lambda_{12}\lambda_{21}-\lambda_{11}\lambda_{22})^2} \end{pmatrix} \tag{5.10}$$

**Proposition 5.4** (Multivariate Gaussian geodesics in mean-precision coordinates)**.**

# 6 Portfolio choice

Having built up the necessary background on information and Riemannian geometries, we are now ready to a concrete economic application. We illustrate our information geometric framework through the classical problem of portfolio choice in a Gaussian setting with myopic CARA utility investors.

Let's describe our economy. Suppose there is a risk free asset with constant instantaneous return $r_f$ and $n$ risky assets. Let $\{R(t)\}_{t\in[0,T]}$ represent the *excess* return *knowledge* of an investor from time $t = 0$ to $t = T$. That is, the $n$-vector $R(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t))$ (i.e. mean $\boldsymbol{\mu}(t)$ and precision $\boldsymbol{\Lambda}(t)$) is an investor's best available information about the instantaneous return of the $n$ risky assets at time $t$. In our economy, all investors are myopic and have CARA preferences with absolute risk aversion parameter $\eta > 0$. In particular, this means investors will treat $R(t)$ as the instantaneous return of the $n$ risky assets at time $t$ when making portfolio allocations. In latter discussions where we have multiple investors with heterogeneous knowledge, the $R(t)$'s could vary for different investors.

Let's relate the return trajectory $\{R(t)\}$ to our developed information acquisition framework. On the one hand, we can simply think of $R(t)$'s as a sequence of arbitrary return beliefs of an investor; this could represent the information of an *uninformed* investor. On the other hand, this return trajectory takes on a far more meaningful interpretation when we incorporate information acquisition framework of Lemma 2.2; this could represent the information of an *informed* investor. At $t = 0$, an investor is endowed with the initial knowledge $R(0) \sim \mathcal{N}(\boldsymbol{\mu}(0), \boldsymbol{\Lambda}(0))$ of the true return $\hat{R}$, and the risk free rate $r_f$ is known knowledge to the investor. The investor endogenously acquires information to get from his initial knowledge to the terminal truth knowledge. Let $\{R(t)\}_{t\in[0,T]}$, with $R(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t))$, be an optimal Bayesian information acquisition trajectory, with $R(T) = \hat{R}$ with $\boldsymbol{\mu}(T) = \hat{\boldsymbol{\mu}}, \boldsymbol{\Lambda}(T) = \hat{\boldsymbol{\Lambda}}$. Observe that each $R(t)$ precisely represents the best available information at time $t$ of the truth return $\hat{R}$, and the investor will use $R(t)$ as the instantaneous return to make portfolio allocations at time $t$. We emphasize that both interpretations of the trajectory $\{R(t)\}$ will play an important role in our equilibrium asset pricing discussions in Section 7. Indeed, the key strength of the information geometry framework is the flexibility in understanding the trajectory $\{R(t)\}$.

Let's summarize the portfolio allocation problem. Since the investor has myopic CARA preferences, we may normalize his wealth to \$1 at each point in time $t$. As discussed above, we can think of the trajectory $\{R(t)\}$ in two different perspectives. The first perspective is that of an uninformed investor, who essentially arbitrarily selects a belief $R(t)$ of the risky asset returns. This uninformed investor has the portfolio allocation problem,

$$\boldsymbol{\pi}(t) = \arg\max_{\tilde{\boldsymbol{\pi}}\in\mathbb{R}^n} \mathbb{E}[-e^{-\eta(r_f+\tilde{\boldsymbol{\pi}}^\top R(t))}] = \frac{1}{\eta}\boldsymbol{\Lambda}(t)\boldsymbol{\mu}(t), \tag{6.1a}$$

subject to:

$$t \mapsto (\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t)) \text{ is a smooth curve in } M\text{, and} \tag{6.1b}$$

$$R(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t)). \tag{6.1c}$$

On the other hand, from the perspective of an informed investor, his portfolio allocation problem is,

$$\boldsymbol{\pi}(t) = \arg\max_{\tilde{\boldsymbol{\pi}}\in\mathbb{R}^n} \mathbb{E}[-e^{-\eta(r_f+\tilde{\boldsymbol{\pi}}^\top R(t))}] = \frac{1}{\eta}\boldsymbol{\Lambda}(t)\boldsymbol{\mu}(t), \tag{6.2a}$$

subject to:

$$p = (\boldsymbol{\mu}(0), \boldsymbol{\Lambda}(0)) \tag{6.2b}$$

$$q = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}}) \tag{6.2c}$$

$$\gamma \text{ solves } d(p, q)\text{, and } \gamma(t) = (\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t)) \tag{6.2d}$$

$$R(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}(t)) \tag{6.2e}$$

where $d$ is the distance function of (4.1).

The investor allocates the following instantaneous amount of wealth into the $n$ risky assets,

$$\boldsymbol{\pi}(t) = \underset{\tilde{\boldsymbol{\pi}} \in \mathbb{R}^n}{\arg\max}\, \mathbb{E}[-e^{-\eta(r_f + \tilde{\boldsymbol{\pi}}^\top R(t))}] = \frac{1}{\eta}\boldsymbol{\Lambda}(t)\boldsymbol{\mu}(t). \tag{6.3}$$

For the remainder of the paper, we will mostly work with mean-precision coordinates. The advantage of mean-precision coordinates (as opposed to, say, mean-variance-covariance coordinates) is that we have a bilinear expression in the optimal portfolio (6.3). But as emphasized throughout the paper, this choice of coordinates to represent knowledge is arbitrary. If we use different coordinates, the analytical form of the portfolio $\boldsymbol{\pi}(t)$ will obviously differ. However, all the subsequent portfolio and asset pricing geometric results are "equivalent" up a diffeomorphism of the coordinate change. Finally as a technical remark, we are heavily taking advantage of the fact that our portfolio map is from the space of probability distributions to the Euclidean space. That is, while the investor's knowledge lies on a statistical manifold, the endogenous actions (e.g. portfolio choice) based on this knowledge belong in the Euclidean space (i.e. an Euclidean manifold with the Euclidean Riemannian metric). There are numerous special properties of the Euclidean space that we heavily exploit in this paper when discussing the investor's endogenous actions.

## 6.1 Portfolio geometry

Our framework naturally lends itself to several curve geometry inspired definitions.

**Definition 6.1** (Portfolio geometry)**.** In mean-precision coordinates and given a risky return trajectory $\{R(t)\}$, the investor's optimal portfolio choice is $\boldsymbol{\pi}(t)$ from (6.3). We define the *portfolio velocity* as $\boldsymbol{v}(t) := \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\pi}(t) = \dot{\boldsymbol{\pi}}(t)$, and the *portfolio acceleration* as $\boldsymbol{a}(t) := \dot{\boldsymbol{v}}(t)$. The *portfolio speed* is $||\boldsymbol{v}(t)||$. Here, $||\cdot||$ is the $\mathbb{R}^n$ Euclidean norm, and $\cdot$ is the Euclidean inner product.

We define the *unit tangent portfolio vector* as $\boldsymbol{T}(t) := \boldsymbol{v}(t)/\,||\boldsymbol{v}(t)||$. The *scalar component of portfolio acceleration along the velocity* is,

$$a_T(t) := \mathrm{comp}_{\boldsymbol{v}}\boldsymbol{a} = \frac{\boldsymbol{v}(t)\cdot\boldsymbol{a}(t)}{||\boldsymbol{v}(t)||}.$$

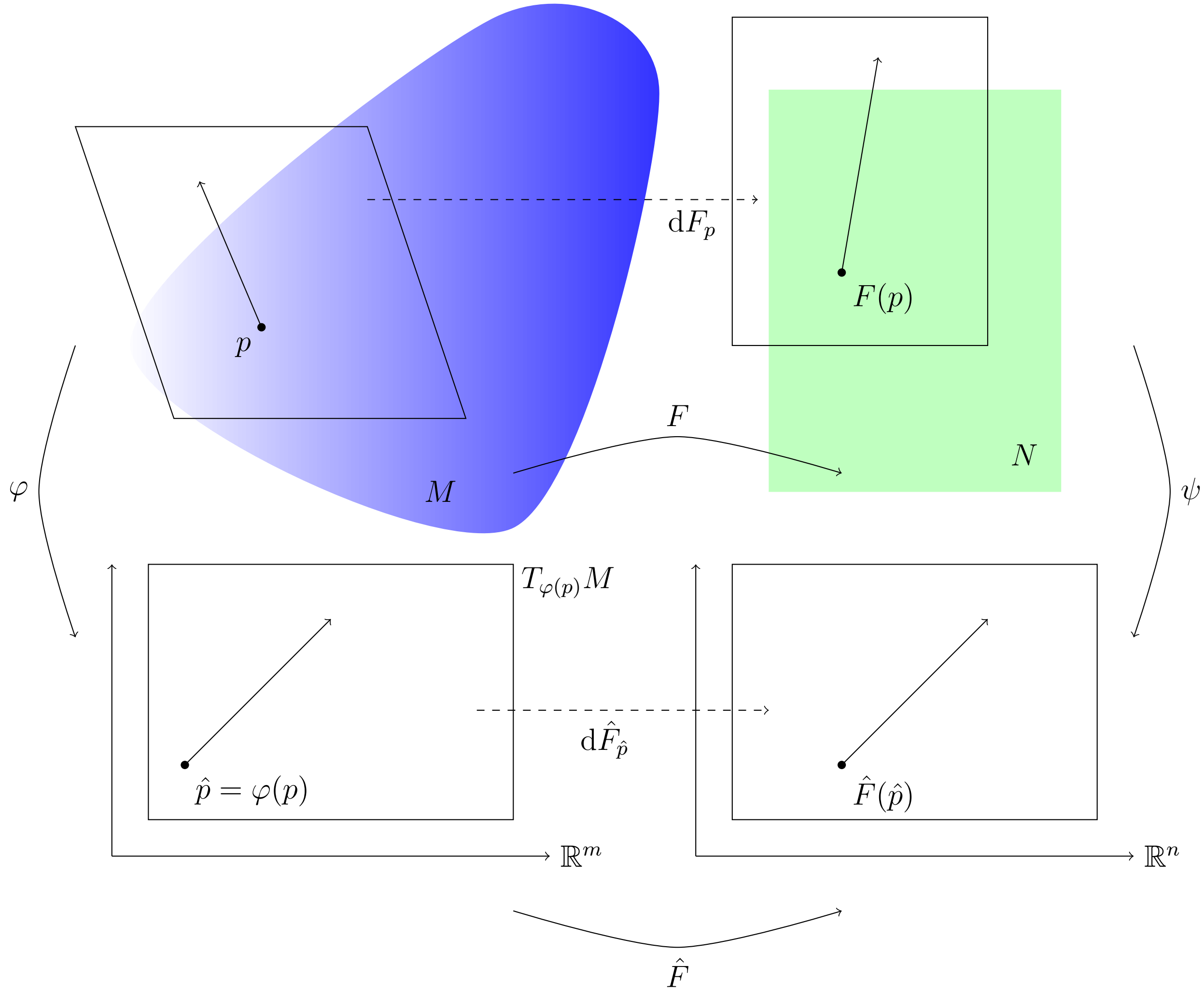

**Figure 4: Map from intrinsic knowledge to portfolio choice.**

The *vector component of portfolio acceleration along the velocity* is,

$$\boldsymbol{a}_T(t) := \mathrm{proj}_{\boldsymbol{v}}\boldsymbol{a} = a_T(t)\,\boldsymbol{T}(t) = \frac{\boldsymbol{v}(t)\cdot\boldsymbol{a}(t)}{||\boldsymbol{v}(t)||^2}\boldsymbol{v}(t).$$

The *vector normal component of portfolio acceleration* is,

$$\boldsymbol{a}_N(t) := \boldsymbol{a}(t) - \boldsymbol{a}_T(t).$$

The *scalar component of portfolio acceleration* is,

$$a_N(t) := \sqrt{||\boldsymbol{a}(t)||^2 - a_T(t)^2} = ||\boldsymbol{a}_N(t)||\,.$$

The following is a standard result in curve geometry.

**Proposition 6.1** (Change in speed, change in acceleration)**.** *Let $\boldsymbol{\pi}(t)$ be a portfolio trajectory. Then,*

*(i) The change in portfolio speed equals the tangential component of the portfolio acceleration, $\frac{\mathrm{d}}{\mathrm{d}t}||\boldsymbol{v}(t)|| = a_T(t)$.*

*(ii) The change in portfolio direction equals the normal component of the portfolio acceleration, $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{T}(t) = \frac{1}{||\boldsymbol{v}(t)||}\boldsymbol{a}_N(t)$.*

## 6.2 Single risky asset

Let's first illustrate our framework with the single risky asset case.

**Proposition 6.2** (Single risky asset portfolio geometry)**.** *Suppose there is only a single $n = 1$ risky asset, and both the mean and the precision are unknown to the investor. Recall the constants $c_0, c_1, c_2$ from Proposition 5.2. Then:*

*(i) The portfolio position in the single risky asset at time $t$ is,*

$$\pi(t) = \frac{\cosh^2\left(c_0 + \frac{t}{\sqrt{2}}\right)\left(2c_2\tanh\left(c_0 + \frac{t}{\sqrt{2}}\right) + c_1\right)}{2c_2^2}$$

*(ii) The portfolio velocity at time t is,*

$$\dot{\pi}(t) = \frac{\frac{1}{2}c_1 \sinh\left(2c_0 + \sqrt{2}t\right) + c_2 \cosh\left(2c_0 + \sqrt{2}t\right)}{\sqrt{2}c_2^2}$$

*(iii) The portfolio acceleration at time t is,*

$$\ddot{\pi}(t) = \frac{2c_2 \sinh\left(2c_0 + \sqrt{2}t\right) + c_1 \cosh\left(2c_0 + \sqrt{2}t\right)}{2c_2^2}$$

*Proof of Proposition 6.2.* The results immediately follow from the closed form solution of Proposition 5.2(iii) case (b), and differentiate. □

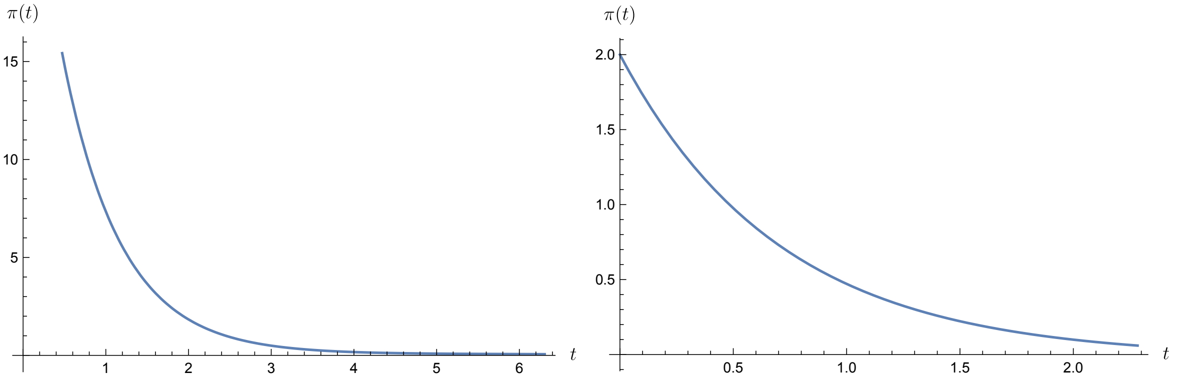

**Figure 5: Portfolio trajectory of the single risky asset.** We illustrate the portfolio trajectory $\pi(t)$ of the single risky asset along the optimal information acquisition trajectory. The left panel illustrates parameter set (1) of Figure 3. The right panel illustrates parameter set (2) of Figure 3.

## 6.3 Multiple risky assets

The case of multiple risky assets is actually far more interesting than the single risky asset case. There are geometric concepts that only have a meaningful definition with multiple risky assets that have no counterpart in the single risky asset case. We note that this phenomenon is an interesting feature of our framework. In classical portfolio choice theory (Markowitz (1952)), and in a gross oversimplification, the extension of a single risky asset portfolio selection to multiple risky asset portfolio selection can be done by rewriting scalar quantities to vector and matrix quantities. The qualitative solution of the classical portfolio choice problem is unchanged from the single risky asset case to the multiple risky asset case. However, as we have discussed in Section 5, dimensionality significantly alters the qualitative and quantitative properties of the information acquisition trajectories. The impact of dimensionality carries over to the portfolio trajectories.

Consider again the portfolio vector $\boldsymbol{\pi}$ of (6.3). The key difference between the multiple risky asset case compared to the single risky asset case is that with more risky assets, we can meaningfully define the ideas of *curvature* and *torsion*. To fix ideas, let's work with $n = 3$ risky assets (as we shall see, the case of $n = 3$ assets is more representative of the general $n$ risky asset case than that of the case of $n = 2$).

Let's construct the *Frenet vectors* associated with the portfolio trajectory $\boldsymbol{\pi}$. Let us define,

$$\begin{aligned}
\boldsymbol{e}_1 &:= \boldsymbol{e}_1(t) = \frac{\dot{\boldsymbol{\pi}}}{||\dot{\boldsymbol{\pi}}||} \\
\boldsymbol{e}_2 &:= \ddot{\boldsymbol{\pi}} - (\ddot{\boldsymbol{\pi}} \cdot \boldsymbol{e}_1)\boldsymbol{e}_1 \\
\boldsymbol{e}_3 &:= \dddot{\boldsymbol{\pi}} - (\dddot{\boldsymbol{\pi}} \cdot \boldsymbol{e}_1)\boldsymbol{e}_1 - (\dddot{\boldsymbol{\pi}} \cdot \boldsymbol{e}_2)\boldsymbol{e}_2
\end{aligned}$$

Observe that $\boldsymbol{e}_1$ is precisely the *unit tangent vector* and $\boldsymbol{e}_2$ is the *unit normal vector*. The vector $\boldsymbol{e}_3$ is the *unit binormal vector* and it is orthogonal to both the unit tangent vector $\boldsymbol{e}_1$ and the unit normal vector $\boldsymbol{e}_2$.

We can define the *portfolio curvature* (*first generalized curvature*) as,

$$\chi_1 := \chi_1(t) = \frac{\dot{\boldsymbol{e}}_1 \cdot \boldsymbol{e}_2}{||\dot{\boldsymbol{\pi}}||}$$

Furthermore, we can define the *portfolio torsion* (*second generalized curvature*) as

$$\chi_2 := \chi_2(t) = \frac{\dot{\boldsymbol{e}}_2 \cdot \boldsymbol{e}_3}{||\dot{\boldsymbol{\pi}}||}.$$

# 7 Asset pricing

The portfolio choice results in Section 6 allow us to discuss several interesting asset pricing implications.

## 7.1 Single risky asset

To fix ideas, let's first work on the case when there is only a single $n = 1$ risky asset in the economy. Let's assume the risky asset is in unit net supply. Then by market clearing, the market price of the risky asset is equal to the aggregate demand of the risky asset. Suppose there are three types of investors in the economy: (i) uninformed of both the mean and precision of the asset return; (ii) informed of the mean but uninformed of the precision; and (iii) uninformed of the mean but informed of the precision. Let's suppose there are, respectively, $w^1, w^2, w^3 > 0$ proportions of these agents in the economy, with $w^1 + w^2 + w^3 = 1$.

Let's make clear on the knowledge set of investor types (i) to (iii). Following Section 6, there is some *truth* distribution $\hat{R} \sim \mathcal{N}(\hat{\mu}, \hat{\lambda})$ of the single risky asset.

The time $t$ *equilibrium price* of the risky asset is,

$$P(t) := \sum_{l=1}^{3} w^k \pi^k(t). \tag{7.1}$$

From this, we can define the *instantaneous time $t$ (log) return* of the risky asset,

$$r(t) := \frac{\mathrm{d}}{\mathrm{d}t} \log P(t) = \frac{\sum_{l=1}^{3} w^k \dot{\pi}^k(t)}{P(t)}. \tag{7.2}$$

Furthermore, we can also define the *velocity* and *acceleration* of returns, respectively, as,

$$v := \dot{r}, \tag{7.3}$$

$$a := \ddot{r}. \tag{7.4}$$

We can further define the *signed curvature* of returns as,

$$k := \frac{a}{(1+v^2)^{3/2}}, \tag{7.5}$$

and also the *curvature* of returns as,

$$\kappa := |k|. \tag{7.6}$$

**Proposition 7.1** (Equilibrium return dynamics of single risky asset)**.** *Suppose there is a representative investor in the economy and in addition to the risk free asset, the single risky asset is in unit net supply. Recall the setup of Proposition 6.2. The equilibrium return $r(t)$ of the single risky asset then satisfies:*

*(i) The time $t$ to $t + \mathrm{d}t$ instantaneous return of the risky asset is,*

$$r(t) = \frac{\operatorname{sech}^2\left(c_0 + \frac{t}{\sqrt{2}}\right)\left(c_1 \sinh\left(2c_0 + \sqrt{2}t\right) + 2c_2 \cosh\left(2c_0 + \sqrt{2}t\right)\right)}{\sqrt{2}\left(2c_2 \tanh\left(c_0 + \frac{t}{\sqrt{2}}\right) + c_1\right)}$$

*(ii) The return velocity is,*

$$v(t) = \frac{1}{2}\left(\operatorname{sech}^2\left(c_0 + \frac{t}{\sqrt{2}}\right) + \frac{c_1^2 - 4c_2^2}{\left(2c_2 \sinh\left(c_0 + \frac{t}{\sqrt{2}}\right) + c_1 \cosh\left(c_0 + \frac{t}{\sqrt{2}}\right)\right)^2}\right)$$

### 7.1.1 Empirical application

Easy and direct empirical applications are possible once we construct the empirical counterpart to those theoretical constructs above. Let's first recall the *finite difference* of a smooth function $f : \mathbb{R} \to \mathbb{R}$. For sufficiently small $h > 0$, the *first order backward difference* approximates the first derivative $f'$,

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}, \tag{7.7}$$

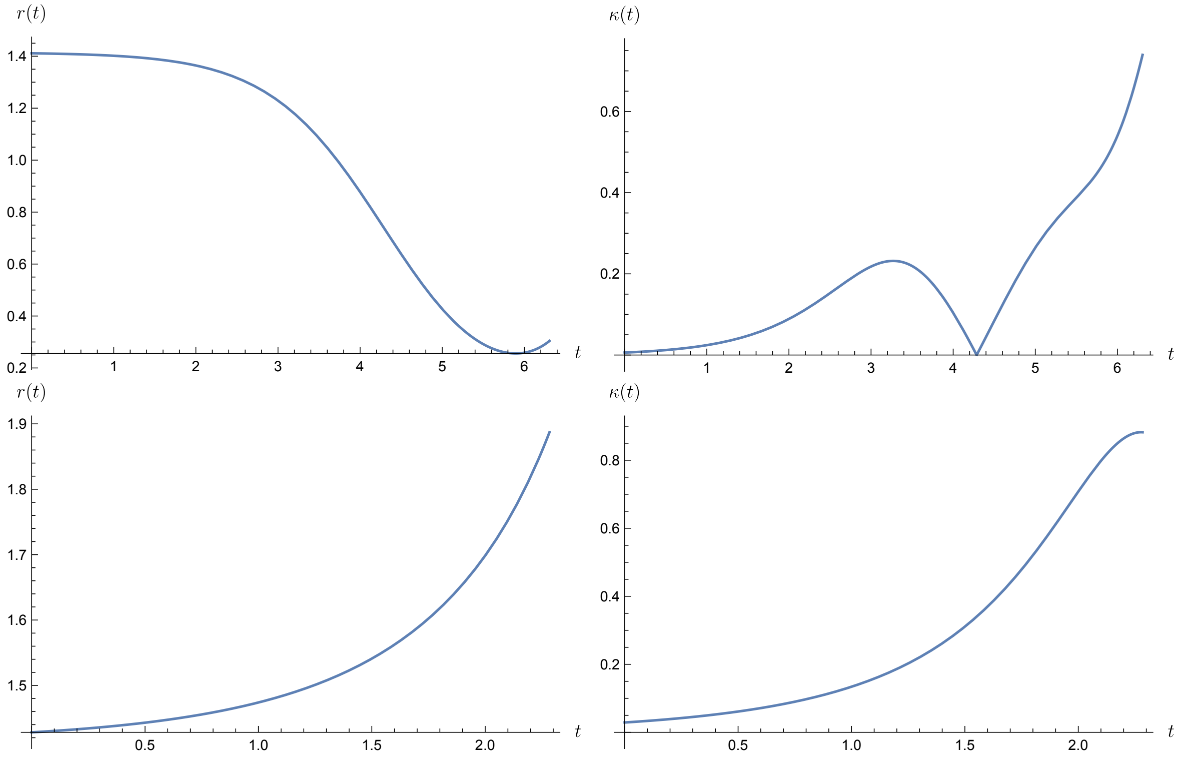

**Figure 6: Illustrations of the single risky asset instantaneous return $r(t)$ and the return absolute curvature $\kappa(t)$.** We illustrate the single risky return $r(t)$ of Proposition 7.1 and the associated return absolute curvature $\kappa(t)$ along the optimal information acquisition trajectory. The top panel illustrates the parameter set (1) of Figure 3. The bottom panel illustrates the parameter set (2) of Figure 3.

while the *second order backward difference* approximates the second derivative $f''$,

$$f''(x) \approx \frac{f(x) - 2f(x-h) + f(x-2h)}{h^2}. \tag{7.8}$$

Suppose we empirically observe returns $r(1), r(2), \ldots, r(T)$ at equidistant increment $h = 1$. Then the empirical counterpart to velocity and acceleration can be constructed by $\hat{v}$ and $\hat{a}$, where

$$\hat{v}(t) := r(t) - r(t-1), \quad t \geq 2, \tag{7.9}$$

$$\hat{a}(t) := r(t) - 2r(t-1) + r(t-2), \quad t \geq 3. \tag{7.10}$$

Analogously, the empirical signed curvature and absolute curvature can be computed as,

$$\hat{k}(t) := \frac{\hat{a}(t)}{(1 + \hat{v}(t)^2)^{3/2}} \tag{7.11}$$

$$\hat{\kappa}(t) := |\hat{k}(t)| \tag{7.12}$$

Observe that there are several differencing schemes for numerically approximating a function (e.g. forward and central differencing), but we deliberately focus on backward differencing. This is so that the empirical quantities only depend on information up to time $t$, and does not have statistical look ahead bias.

## 7.2 Multiple risky assets

The case of multiple risky assets is not a simple technical extension of the single risky asset case. Indeed with multiple assets, we can define several interesting quantities that has no counterpart in the single asset case. Let $\boldsymbol{\pi}^l = (\pi^{l,1}, \ldots, \pi^{l,n})$ denote the portfolio allocation into $n$ risky assets of investor type $l = 1, \ldots, L$.

Let's assume all $n$ risky assets are in unit net supply. By market clearing, the equilibrium price of the $i$-th risky asset is,

$$P^i(t) = \sum_{l=l}^{L} \pi^{l,i}(t),$$

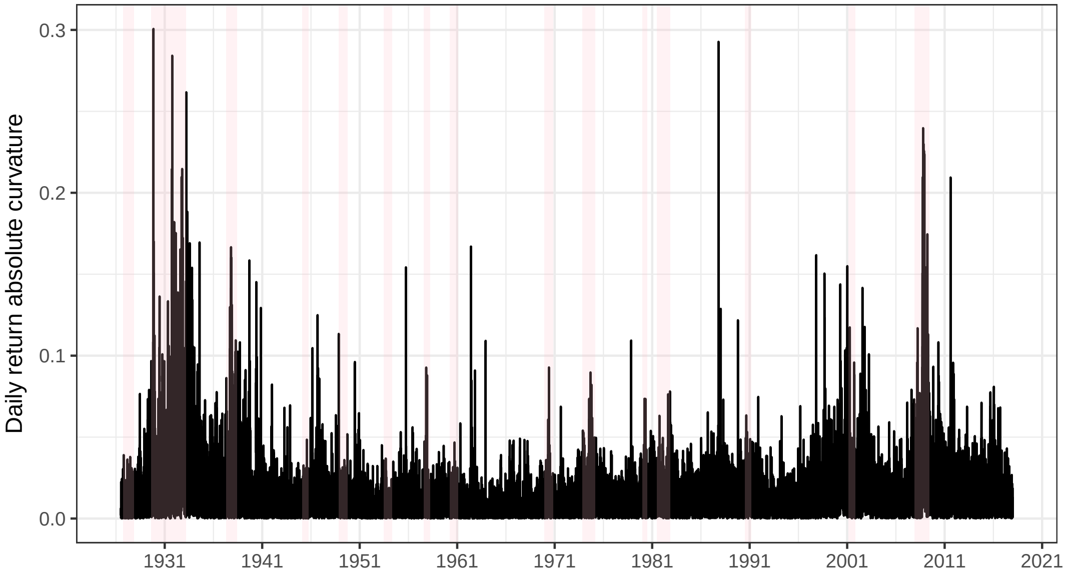

**Figure 7: Daily return absolute curvature of the Fama-French market factor.** The daily return absolute curvature variable is constructed using the daily return data of the Fama-French market factor, and applied to the empirical counterpart to (7.6) using the numerical derivative approximations (7.7) and (7.8). The shaded regions are NBER recession dates.

|  | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| (Intercept) | $15.35^{***}$ | $12.68^{***}$ | $12.79^{***}$ | $13.39^{***}$ | $15.45^{***}$ |
|  | (0.22) | (0.32) | (0.31) | (0.26) | (0.22) |
| return_curv | $215.68^{***}$ | $114.82^{***}$ | $115.39^{***}$ | $62.47^{***}$ | $218.36^{***}$ |
|  | (11.13) | (6.09) | (6.05) | (9.23) | (11.35) |
| lag(return_curv, 1) |  | $66.49^{***}$ | $70.57^{***}$ | $56.51^{***}$ |  |
|  |  | (6.32) | (5.74) | (4.78) |  |
| lag(return_curv, 2) |  | $66.42^{***}$ | $65.82^{***}$ | $45.46^{***}$ |  |
|  |  | (6.42) | (5.59) | (5.34) |  |
| lag(return_curv, 3) |  | $111.28^{***}$ | $109.85^{***}$ | $80.70^{***}$ |  |
|  |  | (6.13) | (6.00) | (6.06) |  |
| mktrf |  |  | $-94.75^{***}$ |  | $-105.43^{***}$ |
|  |  |  | (9.17) |  | (12.55) |
| lag(mktrf, 1) |  |  | $-87.32^{***}$ |  | $-84.31^{***}$ |
|  |  |  | (9.21) |  | (11.91) |
| lag(mktrf, 2) |  |  | $-90.06^{***}$ |  | $-90.29^{***}$ |
|  |  |  | (9.13) |  | (13.29) |
| lag(mktrf, 3) |  |  | $-74.18^{***}$ |  | $-59.91^{***}$ |
|  |  |  | (10.15) |  | (14.38) |
| squared_mktrf |  |  |  | $3811.95^{***}$ |  |
|  |  |  |  | (607.02) |  |
| lag(squared_mktrf, 1) |  |  |  | $3793.88^{***}$ |  |
|  |  |  |  | (413.06) |  |
| lag(squared_mktrf, 2) |  |  |  | $2983.57^{***}$ |  |
|  |  |  |  | (452.58) |  |
| lag(squared_mktrf, 3) |  |  |  | 1035.35 |  |
|  |  |  |  | (621.52) |  |
| $R^2$ | 0.30 | 0.49 | 0.55 | 0.60 | 0.35 |
| Adj. $R^2$ | 0.30 | 0.49 | 0.55 | 0.60 | 0.35 |
| Num. obs. | 7053 | 7050 | 7050 | 7050 | 7050 |
| RMSE | 6.60 | 5.62 | 5.31 | 4.98 | 6.34 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

**Table 1: Regression of daily VIX on return absolute curvature along with other controls.** We linearly regress the daily VIX level onto the *return absolute curvature*. This variable is constructed using the daily return data of the Fama-French market factor, and applied to the empirical counterpart to (7.6) using the numerical derivative approximations (7.7) and (7.8). We also include the raw daily return, absolute daily return, and squared daily return of the Fama-French market factor as control variables. The parentheses show the Newey and West (1987) standard errors with 6 lags.

and like the single asset counterpart, the instantaneous return of the $i$-th risky asset is,

$$r^i(t) := \frac{\mathrm{d}}{\mathrm{d}t} P^i(t).$$

Collecting all the $n$ risky assets together, the return vector is $\boldsymbol{r} = (r^1, \ldots, r^n)$. We can likewise define the *velocity* and *acceleration* of the return vector, respectively, as $\boldsymbol{v} := \dot{\boldsymbol{r}}$ and $\boldsymbol{a} = \dot{\boldsymbol{v}}$.

To fix ideas, let's consider the case of $n = 3$ risky assets (as we shall see, the case of $n = 3$ is far more representative of the general $n$ risky asset case than that of $n = 2$).

# 8 Information recovery

A large empirical literature is built around observing the portfolio holdings of institutional fund managers. The implicit hope of these papers is that by observing the portfolio choices of the institutional fund manager, the econometrician can infer the information set of these supposed informed investors. To the best of our knowledge, few theory papers [14] discuss the recovery of information from the observed actions or prices. Our framework can explicitly attack this problem.

# 9 Generalization to general expected utility maximization problems

# 10 Conclusion

This paper lays out the principles for an intrinsic theory of information acquisition: (P1) an agent's states of knowledge can be described by probability distributions; (P2) information should be intrinsically and not extrinsically measured; and (P3) agent's expected utility maximizing action choices should only depend on intrinsically measured information. We apply tools from the information geometry literature to address principles (P1) and (P2). These geometric concepts have not yet found applications in the economics and finance literature.

[14] The notable exception being Ross (2015).

We illustrate the wide applicability of these tools in a finance context. Notions such as the velocity, acceleration, curvature, and torsion of portfolio holdings and equilibrium returns provide a new perspective in asset pricing. We provide suggestive empirical evidence to show that geometric concepts on returns can shed new light in the empirical asset pricing literature. Finally, our framework is potentially applicable to all expected utility maximization problems, where parameterized random variables can represent the knowledge state of an agent.

# Appendices

## A Overview of information geometry, differential geometry, and Riemannian geometry

This paper does *not* intend to be a treatise on information geometry (whose foundations depend on differential and Riemannian geometry). However, these geometric concepts are not well known in the economics literature, so we need to devote some considerable length for a brief intuitive overview of the key geometric ideas, and delegate the necessary technical details to the Appendix.

### A.1 Differential geometry

We now describe a *manifold* and how it is the key to intrinsically describe information. In our application, we will take $M$ as the statistical manifold, which represents the collection of all possible configurations of information in the economy.

The study of *differential geometry* is the study of smooth manifolds [15] . It will be hopeless to even attempt an in depth description of differential geometry in this paper. The definition of a manifold can be intuitively described in Figure 8.

Simply put, a manifold is a space that is locally flat but globally not. See Figure 8 for an illustration. The most concrete example of a manifold is earth itself. All local inhabitants (e.g. humans, ants, etc.) will physically feel they are living on a flat surface, and yet the earth is actually a round sphere. One of the key defining characteristic of a smooth manifold is that locally, the manifold resembles a flat Euclidean space, but globally it could take on fairly complex shapes. A cartographer can use longitude and latitude coordinates to describe an inhabitant's position on earth. One can equally use the azimuth and elevation to describe the inhabitant's position. Yet the inhabitant simply does not care how a cartographer describes his physical presence. Thus any intrinsic physical description of the inhabitant's behavior on earth should completely be independent of the arbitrary coordinate system. In fact, we have already met this conundrum in Example 1. If we use the Gaussian distribution to represent an agent's information, then it should not matter the form of parameterization an observer uses to describe it.

Now that we have an intuitive understanding of a manifold, we can now describe the concept of a *tangent vector* at a point $p \in M$. See Figure 9 for an illustration of the *tangent space* $T_pM$ at a point $p$. It can be shown that the tangent space $T_pM$ of a point is a vector space of the same dimension as the manifold $M$, and so one can find a basis $\frac{\partial}{\partial x^1}\big|_p, \ldots, \frac{\partial}{\partial x^m}\big|_p$. For notational ease, we will denote $\partial_i|_p := \frac{\partial}{\partial x^i}\big|_p$. Intuitively, a tangent vector $v \in T_pM$ captures the idea of "change" or "direction". As stressed thus far throughout the discussion, our idea of change or direction must be intrinsic and independent of any specific coordinate choice. Indeed, one may view the core building block of differential geometry as this intrinsic formalization of the tangent vectors.

In our application to information acquisition, we will view these tangent vectors as the intrinsic direction of information acquisition. In a particular coordinate choice, we will identity a basis tangent vector on a statistical manifold to the score function $\partial_i|_p := \frac{\partial \log f(Z;p)}{\partial p^i}$. Section 5 discusses these issues in detail.

We remark one important aspect about tangent spaces of general manifolds that may be a source of confusion. On a Euclidean space $\mathbb{R}^m$, it can be shown the tangent space at each point is isomorphic to the $\mathbb{R}^m$ itself. In particular, this means the tangent spaces at all points in $\mathbb{R}^m$ are identical to each other, and so one does *not* need to refer to a "tangent space *at a point*". This is not the case for a general smooth manifold $M$. The tangent space $T_pM$ at point $p \in M$ and a tangent

[15] Actually, one really needs to make a distinction between a *topological manifold* versus a *differential manifold.* Essentially a topological manifold only discusses the shapes of geometric objects, while a differential manifold further endows a calculus structure on said objects. A *Riemannian manifold* provides the notion of distance on geometric objects and is indeed the focus of our paper; we give an overview of Riemannian manifolds in Section A.2.

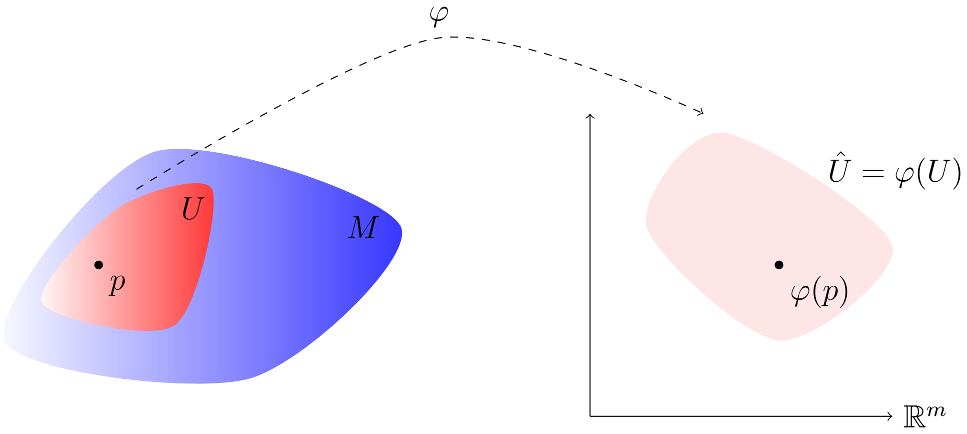

**Figure 8: Smooth manifold.** One of the key defining characteristic of a smooth manifold $M$ is all its small local open sets resemble an $\mathbb{R}^m$ dimensional open subset. A physical analogy is that everything looks flat (e.g. Euclidean) for inhabitants living on the surface of earth, but globally the earth is actually a sphere. Formally, a pair $(U, \varphi)$ is called a *coordinate chart* of a manifold $M$, where $U$ is an open set of $M$ and $\varphi : U \to \hat{U}$ is a $C^\infty$-diffeomorphism called a *smooth coordinate map*, where $\hat{U} = \varphi(U) \subseteq \mathbb{R}^m$. We can represent a point $p \in M$ on the manifold by its *local coordinate representation* $\varphi(p) = (x^1(p), \ldots, x^m(p))$ where $x^i$ is the $i$th component function. For general manifolds, one needs a collection of coordinate charts (called an *atlas*) to cover the entire manifold. In the most general theory, one needs to also discuss how to the overlapping coordinate charts relate to one another. But in information geometry, we are actually endowed with a *global* coordinate chart and hence we do not worry over this technicality here.

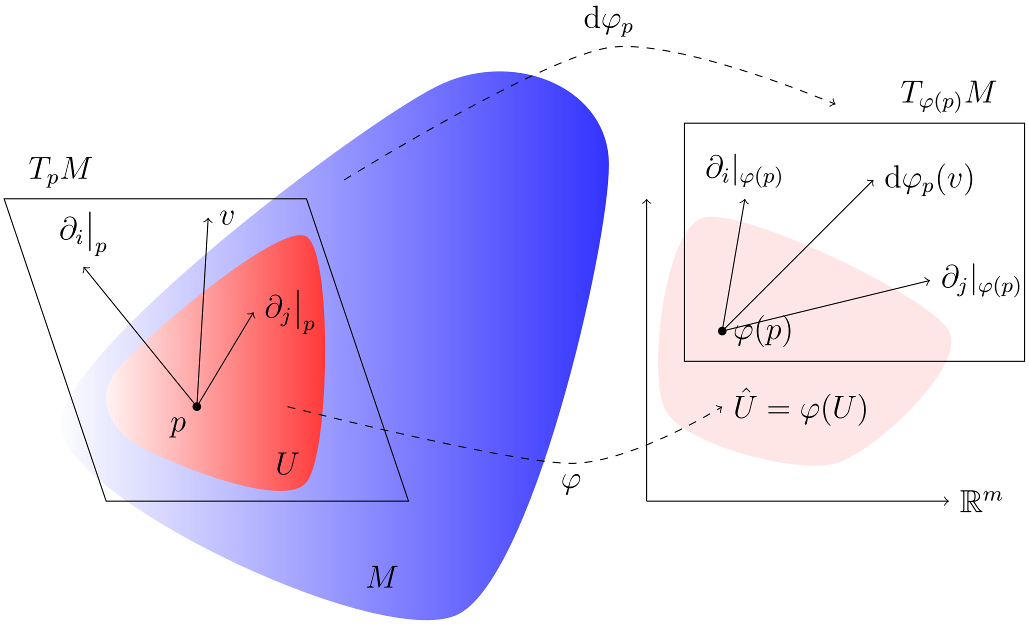

**Figure 9: Tangent space at a point $p$ and its coordinate representation**. Given any point $p \in M$ in a smooth manifold, we can define its *tangent space* $T_pM$ at that point. We will call an element $v \in T_pM$ as a *tangent vector* at the point $p$. Formally, a tangent vector at $p$ is a linear map $v : C^\infty(M) \to \mathbb{R}$ that satisfies a formal product rule: $v(fg) = f(p)vg + g(p)vf$ for all $f, g \in C^\infty(M)$. Furthermore, it is easy to that the tangent space is a vector space of the same dimension as the manifold $M$, and thus we can identity a basis $\frac{\partial}{\partial x^1}\big|_p, \ldots, \frac{\partial}{\partial x^m}\big|_p$. For notational ease, we denote each basis vector by $\partial_i|_p := \frac{\partial}{\partial x^i}\big|_p$ Given a coordinate chart $(U, \varphi)$, we can use the differential $\mathrm{d}\varphi_p$ to represent the tangent space $T_pM$ with coordinates in $T_{\varphi(p)}M$.

space $T_qM$ at point $q \in M$ are in general different. With only a smooth structure on the manifold, it is not straightforward to "compare" these two tangent spaces. The Riemannian manifold structure that we describe next will facilitate this discussion.

## A.2 Riemannian geometry

Now that we have a notion of calculus on a smooth manifold, we now can discuss the idea of "distance" on a manifold. A *Riemannian manifold* is a smooth manifold endowed with a collection of functions [16] $g_p$ defined on its tangent spaces. This $g$ is known as the *Riemannian metric* and takes a symmetric and positive definite form $g_p : T_pM \times T_pM \to \mathbb{R}$ for each point $p \in M$. In particular, with the Riemannian metric $g$ we can make concrete the geometric notions of distance between points, angles, lengths of curves, area, volume and curvature of the manifold. For instance, given two tangent vectors $v, w \in T_pM$, we can define their *inner product* by $\langle v, w \rangle_p := g_p(v, w)$. Likewise, we can define the *length* of a tangent vector by $||v|| := \sqrt{g_p(v, v)}$. Recall that $\partial_1|_p, \ldots, \partial_m|_p$ form a basis for $T_pM$. Applying the Riemannian metric $g_p$ to this basis, we find that

$$g_p = g_{ij}(p)\mathrm{d}x^i\mathrm{d}x^j, \tag{A.1}$$

where $g_{ij}(p) := g(\partial_i|_p, \partial_j|_p)$ and the matrix $[g_{ij}]$ is symmetric, positive definite and smooth in $p$.

For our purposes towards an intrinsic framework of information acquisition, we are most interested with a special type of curves (the "geodesics") defined on the manifold and its relationship to curvature. Furthermore for the concrete economic application of this paper, we are most interested in the Riemannian geometry of both the univariate and multivariate Gaussian distributions of Section 5.

As a simple example, consider the Euclidean space $M = \mathbb{R}^m$. We can identity [17] the basis tangent vector $\partial_i|_p \in T_p\mathbb{R}^m$ with the standard basis vector $e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^m$. The canonical Riemannian metric on a Euclidean space $\mathbb{R}^m$ is $g_{ij}(p) = \langle e_i, e_j \rangle = \delta_{ij}$ for all $p$, and where $\delta_{ij} = 1$ if $i = j$, and 0 if $i \neq j$. We caution that although in the Euclidean space, $\partial_i|_p$ and $\partial_j|_p$ for $i \neq j$ are canonically orthogonal tangent vectors, this is definitively not true for general manifolds, and especially not true for our information geometric setup. The most important Riemannian metric in our context is the Fisher information metric, where $g_{ij}(p)$ is given by (2.1). We will further develop these ideas in Section 5.

## A.3 Geodesics

The concept of *geodesics* is of utmost importance for our intrinsic theory of information acquisition. Simply put, geodesics describe "straight lines" on a curved manifold. In particular, we will use geodesics to model the concept of trajectory of information acquisition, and also the shortest trajectory of information acquisition between some initial knowledge to some terminal truth knowledge. The concept of geodesics will provide the concrete solution to the concept of Bayesian news acquisition trajectory of Lemma 2.2.

There are two equally important perspectives to think about geodesics: (a) the shortest path perspective; and (b) the straight line perspective. To solidify these ideas for our ultimate information acquisition application, it will be useful to take two slight detours and think about simpler real world physics problems.

### A.3.1 How to fly from New York to London?

Let's begin with (a). Suppose an economist wishes to take a flight from New York to London. On a typical (Mercator projection) world map that's printed on a flat piece of paper, the shortest route will be a straight line. After the flight has landed safely in London, the economist will realize that the actual flight trajectory as plotted on the flat world map paper is actually not

[16] Actually $g$ is a 2-tensor field that is positive definite on the manifold $M$. Thus the Riemannian metric is an intrinsic geometric object that is invariant to arbitrary coordinate choices.

[17] We remark that this identification of the tangent space $T_p\mathbb{R}^m$ to its manifold $M = \mathbb{R}^m$ is in general not possible. Indeed, it can be shown that the only manifold that enjoys this property is the Euclidean space.

straight but is curved, and indeed approaches the north pole. Why does the shortest route from New York to London approach the north pole? This can be made concrete by inspecting the Riemannian metric on the sphere.

Let $\theta$ represent the angle around the equator (i.e. longitude), and let $\phi$ represent the angle from the north pole (i.e. similar to latitude). It can be shown the line element is given by $\mathrm{d}s^2 = r^2 \mathrm{d}\phi^2 + r^2 \sin^2(\phi)\mathrm{d}\theta^2$, and where $r$ is the radius of the earth. This is equivalent to saying the Riemannian metric on the sphere is given by $g_{\phi,\phi}(p) = r^2, g_{\theta,\theta}(p) = r^2 \sin^2(\phi), g_{\phi,\theta}(p) = 0$, where $p = (\theta, \phi)$ is the point on earth. In contrast, the usual flat space $\mathbb{R}^2$ the line element is given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$. The length of a trajectory between two points $p$ and $q$ is given by the line integral $\int_p^q \sqrt{\mathrm{d}s^2}\,\mathrm{d}t$. We see that by inspecting the line element $\mathrm{d}s^2$, the length of the trajectory decreases when $\sin(\phi) \to 0$, which is equivalent to $\phi \to 0$ (e.g. north pole) or $\phi \to \pi$ (e.g. south pole). Since New York is located at north of the equator, this discussion shows, indeed, the shortest path to London is to take a *direction* towards the north pole.

The properties of the Riemannian metric $g$ contains significant (actually *all*) information to the minimal trajectory length problem. As we shall see in our information acquisition context, the form of the Riemannian metric will tell us precisely the *optimal direction of information acquisition.*

### A.3.2 Detour: What is a straight line?

The next detour that we need is to think about what "straight lines" actually mean. On a flat Euclidean space, it is intuitively clear that these three notions are equivalent: (a) a straight line between two points; (b) the trajectory with the shortest distance between two points; and (c) a free particle travelling between two points experience zero acceleration on this trajectory. Let's focus on (c) and consider a particle on a flat space with position $\gamma(t) \in \mathbb{R}^m$ at time $t$. From elementary physics, we regard $\dot{\gamma}(t)$ as the *velocity* and $\ddot{\gamma}(t)$ as the *acceleration* of the particle at time $t$. If the particle's equation of motion is given by $\ddot{\gamma}(t) = 0$ so that the particle experiences no acceleration at all times, then its trajectory must satisfy the equation $\gamma(t) = p + vt$, where $p = \gamma(0)$ is the initial position of the particle, and $v = \dot{\gamma}(0)$ is its initial velocity. Using the standard Euclidean metric, it is also immediate that the shortest distance between two points $p, q \in \mathbb{R}^m$ is a straight line. This means if our particle starting at $p = \gamma(0)$ wishes to reach to a point $q = \gamma(T)$ at some terminal time $T$, it simply needs to direct its initial velocity $v = \dot{\gamma}(0)$ in the correct direction at the beginning. This shows that claims (a), (b) and (c) are equivalent.

### A.3.3 Geodesic equation

"Straight lines" in a general manifold are called *geodesics*, and they are conveniently characterized by an analogous no acceleration condition. In a Euclidean space, the no acceleration condition reads as $0 = \frac{\mathrm{d}}{\mathrm{d}t}\dot{\gamma} = \lim_{h\to 0} \frac{\dot{\gamma}(t+h)-\dot{\gamma}(t)}{h}$. Here, the subtraction $\dot{\gamma}(t+h)-\dot{\gamma}(t)$ makes sense because the two tangent vectors live in the same vector space, $\dot{\gamma}(t+h) \in T_{\gamma(t+h)}\mathbb{R}^m \cong \mathbb{R}^m$, and $\dot{\gamma}(t) \in T_{\gamma(t)}\mathbb{R}^m \cong \mathbb{R}^m$. However, as we have observed in Section A.1 and Figure 9, for a general manifold $M$, we know that the two tangent vectors $v \in T_pM$ and $w \in T_qM$ do not live in the same vector space, and hence subtracting $v$ from $w$ does not make sense. In place of the ordinary derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ on Euclidean space, we use instead the *covariant derivative* $D_t$ that, intuitively, "error adjusts" for making computations between two different tangent spaces. The correct formulation of the no acceleration condition using the covariant derivative is $D_t\dot{\gamma} = 0$; this is called the *geodesic equation* and its solution trajectories $\gamma$ are called *geodesics*. See Figure 10 for an illustration.

While the covariant derivative $D_t$ is intrinsically geometrically sound, we still need an explicit coordinate representation of the geodesic equation in order to make concrete computations. Let $(U, \varphi)$ be a coordinate chart of $M$ and let $p \in M$ have the coordinate representation $\varphi(p) = (x^1(p), \ldots, x^m(p)) \in U \subset \mathbb{R}^m$. This means any curve $\tilde{\gamma} : I \to M$, where $I$ is an interval in $\mathbb{R}$, has the coordinate representation $\gamma(t) = \varphi \circ \tilde{\gamma}(t) = (\gamma^1(t), \ldots, \gamma^m(t))$. From the definition of covariant differentiation, one can show that the coordinate representation of the geodesic equation is the system of ordinary differential equations (ODE),

$$0 = \frac{\mathrm{d}^2\gamma^k}{\mathrm{d}t^2} + \Gamma^k_{ij}\frac{\mathrm{d}\gamma^i}{\mathrm{d}t}\frac{\mathrm{d}\gamma^j}{\mathrm{d}t}, \tag{A.2}$$

where $\Gamma^k_{ij}$ are the *Christoffel symbols* of the Riemannian metric $g$. Notice that if $\Gamma^k_{ij} \equiv 0$, then we recover exactly the no acceleration condition for a trajectory on a flat Euclidean space. The Christoffel symbols in coordinates are given by,

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}), \tag{A.3}$$

where $[g^{ij}]$ is the inverse of the matrix $[g_{ij}]$, so $g_{ml} g^{lk} = \delta^k_m$. The Christoffel symbols are the key inputs to make precise the intuitive notion of "curvature" on a manifold.

For our intrinsic information acquisition application, the concept of geodesics is the most important idea to draw from information geometry, differential geometry and Riemannian geometry. However, to even give an intuitive non-rigorous definition of a geodesic, one must have at least some intuitive notions laid out in Sections A.1 and A.2. Indeed, one may view (A.2) as the true starting point of our paper, and everything before it simply lays its groundwork. We will be explicitly concrete in writing down these geodesics in Section 5, and more importantly, explain the significance of those Christoffel symbols in our context of intrinsic information acquisition.

But in order to facilitate a discussion of geodesics, we must first discuss the idea of *parallel transport* of tangent vectors. As we have discussed in Section A.1, given that the tangent space at one point is not identical to the tangent space at another point, a tangent vector $v \in T_p M$ cannot be compared to another tangent vector $w \in T_q M$. An intuitively appealing idea to compare a vector $v \in T_p M$ to $w \in T_q M$ is to "push" or "transport" the vector $v$ to the tangent space $T_q M$. Furthermore, one wishes to transport a vector $v$ in a "parallel" fashion.

## A.4 Curvature

### A.4.1 Jacobi field

The key idea in Section 4.1 is that the Fisher information metric provides a way to canonically measure the distance between random variables. Thus, the Fisher information metric satisfies principle (a). However, the Fisher information metric still inherently depends on the parameterization $\theta$ chosen for the distribution of a random variable. The ideas of Riemannian geometry indeed show that the Fisher information metric is a *Riemannian metric* on the manifold of random variables. This in particular implies that all objects discussed on the manifold are intrinsic to the manifold itself, and does not require an "ambient space", like an Euclidean space [18] .

# References

Amari, S.-i. (2016): *Information Geometry and Its Applications*, Springer Japan.

Amari, S.-i. and H. Nagaoka (2007): *Methods of Information Geometry*, American Mathematical Society.

Ay, N., J. Jost, H. Le, and L. Schwachhöfer (2017): *Information Geometry*, Springer International Publishing.

Bernoulli, J. (1713): *Ars Conjectandi.*

[18] The easiest analogy here is the earth itself. In order to discuss the trajectories of particles on the surface of the earth, it should not depend on whether the earth is embedded in a three dimensional Euclidean space or not.

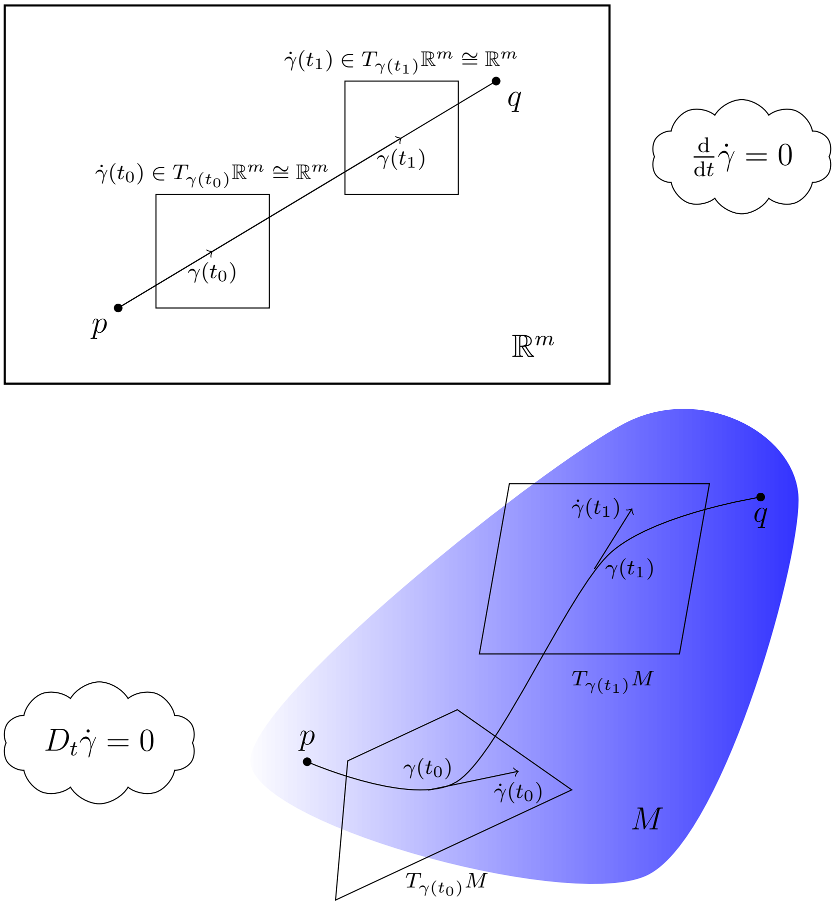
$\dot{\gamma}(t_1) \in T_{\gamma(t_1)}\mathbb{R}^m \cong \mathbb{R}^m$
$q$
$\gamma(t_1)$
$\frac{\mathrm{d}}{\mathrm{d}t}\dot{\gamma} = 0$
$\dot{\gamma}(t_0) \in T_{\gamma(t_0)}\mathbb{R}^m \cong \mathbb{R}^m$
$\gamma(t_0)$
$p$
$\mathbb{R}^m$
$\dot{\gamma}(t_1)$
$q$
$\gamma(t_1)$
$T_{\gamma(t_1)}M$
$D_t\dot{\gamma} = 0$
$p$
$\gamma(t_0)$
$\dot{\gamma}(t_0)$
$M$
$T_{\gamma(t_0)}M$

**Figure 10: Straight lines and geodesics.** We illustrate the idea of a "straight line" both in a flat Euclidean space $\mathbb{R}^m$ (top) and a general Riemannian manifold $M$ (bottom). A straight line in a flat Euclidean space can be described by the zero acceleration equation $\frac{\mathrm{d}}{\mathrm{d}t}\dot{\gamma} = 0$. This is made possible because the derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ to describe the velocity of the curve lies in the same tangent space at all points in the Euclidean space. This is distinctively not the case as already noted in Figure 9 because of the intrinsic curvature of the manifold. To handle the incompatibility between tangent spaces, one replaces the ordinary derivative like $\frac{\mathrm{d}}{\mathrm{d}t}$ on a Euclidean space with a *covariant derivative* $D_t$ that acts on vector fields. The solutions $\gamma$ to the equation $D_t\dot{\gamma} = 0$ are called *geodesics*. Loosely speaking, the covariant derivative provides a way of differentiation across different tangent spaces by "error adjusting" the differences between the tangent spaces due to the presence of the manifold's curvature. This discussion can be formalized using *parallel transports* via the *Levi-Civita connection* $\nabla$ associated with the Riemannian metric $g$.

Calin, O. and C. Udriste (2014): *Geometric Modeling in Probability and Statistics*, Springer International Publishing.

Calvo, M. and J. M. Oller (1990): "A Distance between Multivariate Normal Distributions Based in an Embedding into the Siegel Group," *Journal of Multivariate Analysis*, 35, 223–242.

——— (1991): "An Explicit Solution of Information Geodesic Equations for the Multivariate Normal Model," *Statistics & Decisions*, 9, 119–138.

Debreu, G. (1972): "Smooth preferences," *Econometrica*, 40, 603–615.

Gray, A. (1974): "The volume of a small geodesic ball of a Riemannian manifold," *The Michigan Mathematical Journal*, 20, 329–344.

Grossman, S. J. and J. E. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, 70, 393–408.

Jaynes, E. T. (1978): "Where do we stand on maximum entropy?" in *The Maximum Entropy Formalism*, ed. by R. Levine and M. Tribus, MIT Press.

Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society A*, 186, 453–461.

Keynes, J. M. (1921): *A Treatise on Probability*, MacMillan and Co., London.

Lee, J. M. (1997): *Riemannian Manifolds: An Introduction to Curvature*, Springer.

Ly, A., M. Marsman, J. Verhagen, R. Grasman, and E. Wagenmakers (2017): "A Tutorial on Fisher Information," .

Markowitz, H. M. (1952): "Portfolio selection," *The Journal of Finance*, 7, 77–91.

Marriott, P. and M. Salmon (2000): *Applications of Differential Geometry to Econometrics*, Cambridge University Press, 1 ed.

Newey, W. and K. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Petersen, P. (2016): *Riemannian Geometry*, Springer, 3 ed.

Rao, C. R. (1945): "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, 37, 81–91.

Ross, S. (2015): "The Recovery Theorem," *The Journal of Finance*, 70, 615–648.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Skovgaard, L. T. (1984): "A Riemannian Geometry of the Multivariate Normal Model," *Scandinavian Journal of Statistics*, 11, 211–223.