Mutual Information Gradient Induces Contraction of Information Radius

Tsutomu Mori*

Department of Human Lifesciences,
Fukushima Medical University School of Nursing, Fukushima, Japan.
**Correspondence:* tmori@fmu.ac.jp

Abstract

Mutual information (MI) quantifies how knowledge of one system reduces uncertainty about another. We show that when interactions between two systems follow the gradient ascent of mutual information, the information radius

$$r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t;Y_t)$$

monotonically contracts along the trajectory. Contraction is exact in the fixed–marginal regime, satisfying $\dot{r}(t) = -2\,\dot{\mathrm{MI}}(t)$, and robust in the small marginal drift regime, where combined marginal entropy drift is dominated by MI gains. This theorem applies to both discrete and continuous settings; for the latter, we employ a copula-based formulation that preserves marginal structure and supports a projected-gradient interpretation. Under isothermal, closed, and quasi-static conditions, the relationship $\Delta U = -k_B T\,\Delta\mathrm{MI}$ emerges, suggesting a mechanical analogy. Beyond this regime, we formalize an Attractive Influential Force (AIF) framework, characterizing informational interactions that guarantee distance contraction. We further demonstrate that exponential kernels based on r and their generalizations using a unified information distance D_{info} share a syntactic invariance: all align qualitatively with the MI gradient. This approach unifies geometric, thermodynamic, and kernel-based perspectives on informational coupling, laying the groundwork for applications in learning dynamics, biological information flow, and networked multi-agent systems.

Keywords: mutual information; information geometry; distance contraction; copula representations; influential forces; informational dynamics

1 Introduction

Mutual information (MI) is one of the most fundamental concepts in information theory [1, 2, 3]. It quantifies the extent to which knowledge of one random variable reduces the uncertainty of another, thereby expressing the informational dependency between systems. While entropy measures uncertainty within a given system, MI characterizes the proportion of structure shared between systems and has traditionally been treated as a scalar measure of dependency, capacity, or predictive power. However, MI encompasses more than a mere numerical summary: it also possesses a *gradient structure*, reflecting how informational relationships evolve as systems interact [4].

This work places that gradient structure at the core of its analytical framework. We show that under mild thermodynamic conditions—namely an isothermal, closed, and quasi-static regime—the gradient of mutual information induces a contraction of the information radius

$$r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t; Y_t), \tag{1}$$

which represents the portion of uncertainty not shared between two evolving systems. As MI increases, this unshared portion necessarily decreases, and the distance $r(t)$ contracts. This yields a geometric view of how informational coupling develops over time and forms the basis of the main result of this paper [5].

Connections between physics and information theory have long been recognized [6, 7, 8]. In statistical thermodynamics, entropy gradients drive dissipative processes, while information geometry introduces Riemannian and divergence-based structures (such as the Fisher–Rao metric and KL-type divergences) to describe distances on manifolds of probability distributions [9]. However, these classical frameworks focus on *differences between distributions*, rather than on *the evolving relationships between interacting systems*. In contrast, our approach addresses this evolution explicitly by introducing an information radius that captures how interdependence changes over time.

This intuitive picture admits a rigorous formulation. As two systems exchange information, their predictive capabilities increase and the information radius between them decreases. When marginal entropies vary slowly or remain constant, increases in MI correspond directly to convergence in information space. Thus the gradient $\nabla\mathrm{MI}$ can be regarded as a natural vector field guiding the systems toward stronger mutual predictability. Theorem 3.1 formalizes this principle, establishing that motion aligned with $\nabla\mathrm{MI}$ monotonically decreases $r(t)$ under broad conditions.

Before introducing the thermodynamic correspondence, it is important to articulate its limitations. Within the isothermal, closed, and quasi-static regime described above, fluctuations in shared information can be associated with changes in internal energy [10, 11, 12]. This suggests a correspondence between informational flow and physical quantities. In this constrained environment, the mechanical manifestation of informational coupling takes the form

$$\boldsymbol{F}_{att} = k_{\mathrm{B}}T\,\nabla\mathrm{MI}, \qquad \Delta U = -\,k_{\mathrm{B}}T\,\Delta\mathrm{MI}, \tag{2}$$

where k_{B} is Boltzmann's constant and T is temperature.

These relations offer two complementary perspectives. First, the MI gradient establishes a directional tendency that moves the systems toward states of higher mutual predictability. Second, increases in shared information reduce internal energy by an amount proportional to $k_{\mathrm{B}}T$, providing an energetic interpretation [8, 13]. Taken together, these expressions demonstrate that informational interactions may adopt the mathematical form of forces, while their interpretation here remains conceptual rather than mechanical.

The implications of this viewpoint are broad. Because MI is defined for random variables of any origin, the contraction phenomenon applies to a wide variety of systems, including physical processes, biological interactions, computational dynamics, and organizational or social structures [14]. The analysis relies solely on MI and its gradient, giving the framework considerable universality and domain independence.

While the present paper is self-contained, it also forms part of a broader program interpreting natural forces as informational interactions, referred to as *Influential Force Theory* (IFT) [**?**]. Within this program, the distance $r(t)$ represents a specific instance of a more general unified information distance D_{info}, whose structural role is examined in later sections [16, 17, 18].

Outline. Section 2 introduces notation, the information radius $r(t)$, and the thermodynamic assumptions. Section 3 states the contraction theorem. Section 4 develops supporting propositions, including the fixed-marginal, small-drift, discrete-step, and noise-tolerant formulations. Section 5 presents geometric interpretations. Section 6 provides examples and applications. Section 7 discusses related work. Section 8 interprets MI gradients as influential forces. Section 9 outlines broader implications and future research directions, and Section 10 concludes the paper.

2 Preliminaries

2.1 Notation and Setup

Let X and Y denote random variables taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint distribution $P_{XY}(x, y)$ and marginals $P_X(x)$ and $P_Y(y)$. The Shannon entropy is

$$H(X) = -\sum_x P_X(x) \log P_X(x), \tag{3}$$

and the mutual information (MI) is defined as

$$\mathrm{MI}(X;Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}. \tag{4}$$

All information-theoretic quantities are measured in natural units (nats) unless stated otherwise [1, 2, 19].

Gradients. When gradients of MI appear later, the notation $\nabla\mathrm{MI}$ may refer either to (i) a gradient with respect to finite-dimensional parameters that govern P_{XY}, or (ii) a functional gradient on the space of joint densities. Only the ascent property—namely that motion is locally aligned with $\nabla\mathrm{MI}$—is required for the main theorem [4, 20].

2.2 Information Radius

To quantify the informational separation between two systems, we consider the symmetric functional

$$r(X, Y) = H(X) + H(Y) - 2\,\mathrm{MI}(X;Y), \tag{5}$$

which represents the uncertainty not shared between X and Y. Using standard identities,

$$r(X, Y) = H(X|Y) + H(Y|X), \tag{6}$$

which makes the interpretation of r as "unshared uncertainty" explicit in the discrete case [2].

Motivation. Classical distances such as KL divergence or Fisher–Rao geometry measure dissimilarities between *fixed* distributions [7]. In contrast, $r(X, Y)$ characterizes a *dynamic* informational relationship between two interacting systems. As $\mathrm{MI}(X;Y)$ increases, the unexplained portions of X and Y decrease, and therefore r contracts [23]. Thus r provides a natural geometric indicator of how strongly two systems become coupled through their informational interaction.

Remark on unified distances. Although the present paper focuses on r, later sections introduce a more general unified information distance D_{info}, which serves as a broader geometric framework for influential forces [16]. Only its structural role is required here; full details appear in Section 7.

Continuous variables. For continuous random variables, differential entropies may be negative and the inequality $\mathrm{MI}(X;Y) \leq \min\{H(X), H(Y)\}$ need not hold. In Sec. 4.3, we therefore adopt a copula-based representation that preserves marginals and expresses MI as an integral of $c(u, v) \log c(u, v)$ over $[0, 1]^2$ [24, 25, 26]. All contraction arguments used later rely only on changes in MI, not on absolute entropy values.

2.3 Thermodynamic Setting

To model the evolution of informational coupling, we adopt a minimal thermodynamic perspective. Throughout this work, we consider the interaction under *mild thermodynamic conditions*—that is, in an isothermal, closed, and quasi-static setting [10, 11, 12]. These assumptions isolate internal informational interactions and ensure that MI and entropy remain well-defined along the trajectory.

Motivation for these conditions. The above assumptions parallel those in equilibrium thermodynamics and allow MI to behave analogously to negative free energy. They provide a clean interpretation of $\nabla\mathrm{MI}$ as a force-like quantity that promotes increased predictability between the interacting systems [8, 13].

Mechanical attractive form. Under the isothermal, closed, quasi-static regime,

$$\boldsymbol{F}_{att} = k_B T \, \nabla \mathrm{MI}, \qquad \Delta U = -\, k_B T \, \Delta \mathrm{MI}. \tag{7}$$

Therefore, an increase in shared information corresponds to a decrease in internal energy [6].

Regimes used later.

- **Fixed marginals:** If $H(X)$ and $H(Y)$ remain constant, then $\dot{r}(t) = -2\,\dot{\mathrm{MI}}(t)$, and any ascent in MI yields exact contraction (see Sec. 4.3) [17, 18].
- **Small marginal drift:** If $|\Delta H(X)| + |\Delta H(Y)| \leq \alpha\,\Delta\mathrm{MI}$ for $0 \leq \alpha < 2$, then $\Delta r \leq (\alpha - 2)\Delta\mathrm{MI} < 0$ (see Sec. 4.2) [27, 28].

Terminology. From this point onward, we refer to any interaction driven by $\nabla\mathrm{MI}$ as an influential force. The mechanical form (7) is treated as an approximative case valid under the thermodynamic assumptions above, while later sections develop the broader Attractive Influential Force used in the IFT framework [**?**].

3 Main Result

To derive the contraction property rigorously, we first specify the informational dynamics that govern the evolution of (X_t, Y_t).

Assumption 1 (MI-gradient dynamics with fixed marginals)**.** *Let* (X_t, Y_t) *evolve according to the mutual-information gradient flow*

$$\frac{dX_t}{dt} = \eta \, \nabla_X \mathrm{MI}(X_t; Y_t), \qquad \frac{dY_t}{dt} = \eta \, \nabla_Y \mathrm{MI}(X_t; Y_t),$$

with $\eta > 0$. *Assume that the marginal entropies* $H(X_t)$ *and* $H(Y_t)$ *remain constant in* t*, corresponding to an isothermal, closed, quasi-static regime in which the internal variability is negligible compared with the variation of the mutual information. We also assume that* $\mathrm{MI}(X_t; Y_t)$ *is differentiable in* t*.*

Theorem 3.1 (Contraction of Information Radius)**.** *Under Assumption 1, the information radius*

$$r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t; Y_t)$$

is monotonically non-increasing along the MI-gradient flow. In particular,

$$\frac{dr(t)}{dt} \leq 0.$$

Proof. The proof follows from the differential identity for $r(t)$ and the monotonicity of MI under the MI-gradient dynamics. □

Connection to Appendix A. A fully rigorous formulation of the MI-gradient flow, including the projected-gradient geometry on the space of joint couplings with fixed marginals, is provided in Appendix A. The simplified presentation in Section 3 offers the intuitive information-theoretic expression of the contraction phenomenon, whereas Appendix A develops the complete variational structure and proves the monotonicity of $\mathrm{MI}(P(t))$ in full mathematical detail. This variational formulation replaces the heuristic state-space dynamics of Assumption 1 with a projected-gradient flow on the coupling space, which rigorously enforces the fixed-marginal constraint rather than merely assuming it.

4 Proof Sketch and Supporting Lemmas

In this section we provide a proof sketch of Theorem 3.1, together with several supporting lemmas that clarify the differential structure underlying the MI-gradient flow. The full variational formulation is developed in Appendix Section A, which follows the classical structure of variational gradient flows and projected couplings [16, 17, 18].

4.1 Differential Identity for Information Radius

Lemma 4.1 (Time derivative of information radius)**.** *Let*

$$r(t) := H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t; Y_t),$$

where each term is differentiable in t*. Then*

$$\frac{dr(t)}{dt} = \frac{dH(X_t)}{dt} + \frac{dH(Y_t)}{dt} - 2\,\frac{d}{dt}\mathrm{MI}(X_t; Y_t).$$

Proof. Differentiate the definition of $r(t)$; this identity is standard in information-theoretic treatments of conditional entropy and mutual information [2]. □

4.2 MI-Gradient Dynamics Increase MI

Lemma 4.2 (MI gradient increases mutual information)**.** *Assume the heuristic MI-gradient dynamics*

$$\frac{dX_t}{dt} = \eta\, \nabla_X \mathrm{MI}(X_t; Y_t), \qquad \frac{dY_t}{dt} = \eta\, \nabla_Y \mathrm{MI}(X_t; Y_t),$$

for some constant $\eta > 0$*, and suppose* $\mathrm{MI}(X_t; Y_t)$ *is continuously differentiable in* t*. Then*

$$\frac{d}{dt}\mathrm{MI}(X_t; Y_t) = \eta\big(\|\nabla_X \mathrm{MI}(X_t; Y_t)\|^2 + \|\nabla_Y \mathrm{MI}(X_t; Y_t)\|^2\big) \;\geq\; 0.$$

Thus MI is non-decreasing along this flow.

Proof. Apply the chain rule and substitute the assumed dynamics; similar arguments appear in classical analyses of gradient-based learning and information-geometric optimization [20]. □

4.3 Proof Sketch of Theorem 3.1

Under Assumption 1, the marginal entropies $H(X_t)$ and $H(Y_t)$ are constant in t. By Lemma 4.1,

$$\frac{dr(t)}{dt} = -\,2\,\frac{d}{dt}\mathrm{MI}(X_t; Y_t).$$

Furthermore, Lemma 4.2 ensures that $\frac{d}{dt}\mathrm{MI}(X_t; Y_t) \geq 0$, and therefore $\frac{dr(t)}{dt} \leq 0$.

$\frac{d}{dt}\mathrm{MI}(X_t; Y_t) \geq 0$, hence $\frac{dr(t)}{dt} \leq 0$. This establishes the contraction property at the heuristic level. The rigorous variational version, formulated on the coupling space with fixed marginals and using projected gradient flows, is developed in Appendix Section A [16, 17].

4.4 Approximate Contraction under Small Marginal Drift

We record a quantitative version of information-distance contraction when marginal drifts are small but not strictly zero.

Proposition 4.3 (Approximate contraction under small marginal drift)**.** *Consider a transition* $(X, Y) \mapsto (X', Y')$ *with changes* $\Delta H(X)$*,* $\Delta H(Y)$*, and* $\Delta\mathrm{MI} = \mathrm{MI}(X'; Y') - \mathrm{MI}(X; Y)$*. Assume* $\Delta\mathrm{MI} > 0$ *and that marginal drift is dominated by MI change:*

$$|\Delta H(X)| + |\Delta H(Y)| \;\leq\; \alpha\, \Delta\mathrm{MI}, \qquad 0 \leq \alpha < 2.$$

Then

$$\Delta r = \Delta H(X) + \Delta H(Y) - 2\,\Delta\mathrm{MI} \;\leq\; (\alpha - 2)\,\Delta\mathrm{MI} \;<\; 0.$$

In particular, for α *small (near fixed marginals),* $\Delta r \approx -2\,\Delta\mathrm{MI}$*.*

Remark 4.4. Defining $E := \Delta r + 2\,\Delta\mathrm{MI} = \Delta H(X) + \Delta H(Y)$, if $|\Delta H(X)| + |\Delta H(Y)| \leq \varepsilon\,|\Delta\mathrm{MI}|$ with $0 \leq \varepsilon < 2$, then $|\Delta r + 2\,\Delta\mathrm{MI}| \leq (\varepsilon/2) \cdot 2|\Delta\mathrm{MI}|$. This type of inequality has analogues in MI estimation theory and nonparametric information bounds [27, 28].

Corollary 4.5 (Differential version)**.** *If* $H(X_t), H(Y_t), \mathrm{MI}(X_t; Y_t)$ *are differentiable and*

$$|\dot{H}(X_t)| + |\dot{H}(Y_t)| \;\leq\; \alpha\, \dot{\mathrm{MI}}(t), \qquad \dot{\mathrm{MI}}(t) > 0,$$

then

$$\dot{r}(t) = \dot{H}(X_t) + \dot{H}(Y_t) - 2\,\dot{\mathrm{MI}}(t) \;\leq\; (\alpha - 2)\,\dot{\mathrm{MI}}(t) \;<\; 0.$$

4.5 Marginal-Preserving Informational Flows

We now describe informational flows that preserve the marginal distributions—relevant for the rigorous formulation in Appendix Section A.

Copula notation. For positive densities $f_X, f_Y, f_{X,Y}$, define

$$u = F_X(x), \qquad v = F_Y(y), \qquad c(u,v) = \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}.$$

Then

$$\mathrm{MI}(X;Y) = \iint c(u,v)\,\log c(u,v)\,du\,dv,$$

the classical copula-based representation of dependence structure [24, 25, 26].

Assumption 2 (Smooth copula path with preserved marginals). *Let $s \mapsto p_s(x,y)$ be a smooth path with copula density $c_s(u,v)$, and assume*

$$\int \dot{p}_s(x,y)\,dy = 0, \qquad \int \dot{p}_s(x,y)\,dx = 0,$$

so that the marginals remain fixed along the path.

Lemma 4.6 (Copula-only generators). *If the marginals are preserved, then for some generator $\dot{c}_s$,*

$$\dot{p}_s(x,y) = f_X(x) f_Y(y)\,\dot{c}_s(u,v),$$

with constraints

$$\int_0^1 \dot{c}_s(u,v)\,du = 0, \qquad \int_0^1 \dot{c}_s(u,v)\,dv = 0.$$

Moreover,

$$\frac{d}{ds}\mathrm{MI}_s = \iint \dot{c}_s(u,v)\,\log c_s(u,v)\,du\,dv.$$

Remark 4.7. These constraints characterize the tangent space of the copula manifold, the natural geometric setting for dependence structures [25].

Proposition 4.8 (Projected gradient on the copula manifold). *If the mechanical Attractive Influential Force induces a projected gradient flow*

$$\dot{c}_s = \Pi_{\mathcal{T}_{c_s}}\big(\nabla_c \mathrm{MI}[c_s]\big),$$

then

$$\frac{d}{ds}\mathrm{MI}_s = \iint \dot{c}_s(u,v)\,\log c_s(u,v)\,du\,dv \;\geq\; 0.$$

Thus MI *increases while $H(X)$ and $H(Y)$ remain constant, consistent with the classical monotonicity results for projected entropy flows and Frobenius-gradient structures [16, 17].*

Corollary 4.9 (Contraction in the copula manifold). *Since $H(X)$ and $H(Y)$ remain constant,*

$$\frac{d}{ds}r(s) = -2\,\frac{d}{ds}\mathrm{MI}(s),$$

hence MI ascent implies contraction of r.

4.6 Exact Contraction under Fixed Marginals

Proposition 4.10 (Exact contraction with fixed marginals)**.** *If $H(X_t)$ and $H(Y_t)$ remain constant and $\dot{\mathrm{MI}}(t) \geq 0$, then*

$$\dot{r}(t) = -2\,\dot{\mathrm{MI}}(t) \;\leq\; 0.$$

If $\dot{\mathrm{MI}}(t) > 0$ on a set of positive measure, then $r(t)$ strictly decreases there. Moreover,

$$r(t_2) - r(t_1) = -2\int_{t_1}^{t_2} \dot{\mathrm{MI}}(t)\,dt \;\leq\; 0.$$

5 Geometric Interpretation

The contraction of information radius admits a compelling geometric interpretation. Although the main result was established analytically in Sections 3 and 4, its conceptual meaning is most transparent when framed in the language of geometric flows on an abstract information space [7, 9].

In this space, each point represents a probability distribution (or informational state), and distances quantify the degree of independence, redundancy, or coupling between components.

5.1 The Information Space and Its Metric Structure

The information radius

$$r(X,Y) = H(X) + H(Y) - 2\,\mathrm{MI}(X;Y)$$

acts as a natural measure of separation between two informational states. Unlike classical geometric distances, r is not derived from a fixed Riemannian metric but instead reflects the structure of statistical dependence. Small values of r correspond to strong coupling (large MI), while large values reflect weaker dependence. This "dependence geometry" viewpoint parallels relational structures in copula theory [24, 25], where dependence is encoded independently of marginal behavior.

In this sense, the pair (X,Y) may be viewed as residing in a manifold of joint distributions, and r behaves like a squared distance along this space. The MI-gradient direction identifies the "steepest ascent'' direction of dependence, and the system moves along curves of increasing mutual information, echoing the structure of information-geometric gradient flows [7].

5.2 MI-Gradient Flow as a Contractive Vector Field

The vector field $\nabla\mathrm{MI}$ generates a flow that decreases the information radius:

$$\frac{dr(t)}{dt} \;\leq\; 0.$$

Thus, MI-gradient dynamics define a *contractive flow* on information radius. This is analogous to geometric flows where curvature induces shrinkage, such as mean-curvature flow or Ricci flow [16, 17]. Here, MI plays the role of an effective scalar potential whose gradient determines the natural direction of evolution.

The flow therefore acts not only analytically but geometrically: it shortens the distance between states in the informational manifold.

5.3 Curvature and Effective Geometry

The persistent decrease of $r(t)$ suggests the presence of an effective positive curvature induced by the MI landscape. While the present work does not commit to a specific geometric structure, the observed contraction parallels behaviors found in classical geometric flows:

- **Mean-curvature flow:** where hypersurfaces contract under their curvature.
- **Ricci flow:** where distances contract in directions of positive curvature.

These phenomena provide a useful analogy, and the MI-gradient flow plays a similar role as a dependence-induced curvature mechanism, guiding trajectories toward regions of stronger coupling [9].

5.4 Potential Landscape Interpretation

Another useful viewpoint is that MI defines a potential landscape over the space of couplings. Regions of high mutual information behave as wells or basins of attraction. Following the MI-gradient flow corresponds to descending this landscape and entering regions where variables are more tightly coupled.

The contraction theorem states that as the system moves downhill in this landscape (i.e., as MI increases), the information radius necessarily shortens. This aligns with the classical idea that information reflects structure, and increasing structure corresponds to geometric convergence [9].

5.5 Relation to Copula Geometry

The copula-based formulation in Section 4 demonstrates that informational dependence can often be understood purely in terms of the dependence structure, independent of the marginals. In copula space, the MI-gradient flow becomes a projected gradient flow on the manifold of valid copulas, and the contraction of r is mirrored within that manifold [25, 26]. This indicates that the contraction property is not tied to a specific coordinate system but is intrinsic to the geometry of dependence itself.

5.6 Connection to the Rigorous Formulation

The geometric interpretation presented here is heuristic but informative. Appendix A provides the corresponding rigorous formulation of the MI-gradient flow as a projected gradient flow on the coupling space with fixed marginals. There, contraction is proved through a variational framework without appealing to any explicit manifold structure [16, 17]. Consequently, the geometric interpretation complements—rather than replaces—the analytic arguments, highlighting the unifying role of mutual information as:

- a scalar potential,
- a geometric functional guiding evolution, and
- a mechanism that shortens information radii.

Thus, the contraction theorem can be viewed simultaneously as an analytic fact and as a geometric phenomenon in the emerging theory of informational dynamics.

6 Examples and Applications

The contraction of the information radius admits a variety of concrete illustrations across discrete, continuous, and more complex statistical models. In this section we present several representative examples and outline applications in physical, biological, and abstract information systems.

6.1 Binary Symmetric Channel (BSC) Toy Model

Consider the simplest setting $X \sim \text{Bernoulli}(1/2)$ and $Y = X \oplus N$ with $N \sim \text{Bernoulli}(p)$ independent. Then $Y \sim \text{Bernoulli}(1/2)$ and

$$\text{MI}(X;Y) = \ln 2 - H_b(p), \qquad H_b(p) := -p\ln p - (1-p)\ln(1-p).$$

The information radius reduces to

$$r(X,Y) = H(X) + H(Y) - 2\,\text{MI}(X;Y) = 2\,H_b(p),$$

consistent with classical channel-capacity analysis in information theory [1, 2]. Differentiating gives

$$\frac{dr}{dp} = 2\,\frac{d}{dp}H_b(p) = 2\ln\frac{1-p}{p},$$

which is positive for $p < 1/2$ and negative for $p > 1/2$. Thus any flow that decreases p (increasing channel reliability) strictly decreases r, in agreement with the MI-gradient intuition.

This example provides a minimal illustration of the contraction phenomenon: as mutual information increases, the information radius r necessarily shrinks.

6.2 Gaussian Pair with Fixed Marginals

Let (X, Y) be zero-mean jointly Gaussian with variances σ_X^2, σ_Y^2 fixed and correlation $\rho \in (-1, 1)$. Then

$$\text{MI}(X;Y) = -\tfrac{1}{2}\ln(1-\rho^2), \qquad H(X) = \tfrac{1}{2}\ln(2\pi e\,\sigma_X^2), \qquad H(Y) = \tfrac{1}{2}\ln(2\pi e\,\sigma_Y^2),$$

so that the information radius becomes

$$r(X,Y) = \tfrac{1}{2}\ln(2\pi e\,\sigma_X^2) + \tfrac{1}{2}\ln(2\pi e\,\sigma_Y^2) + \ln(1-\rho^2),$$

which is a standard result in information geometry and multivariate analysis [7, 9]. Differentiating with respect to ρ gives

$$\frac{\partial r}{\partial \rho} = \frac{d}{d\rho}\ln(1-\rho^2) = -\frac{2\rho}{1-\rho^2},$$

which is strictly negative for $\rho > 0$ and strictly positive for $\rho < 0$. Increasing $|\rho|$ therefore shortens r, illustrating exact contraction.

This is the continuous analogue of the BSC example.

6.3 Information Flow in Markov Chains

Consider a Markov chain $X \to Z \to Y$ with joint distribution P_{XYZ}. Suppose that the intermediate channel $P_{Y|Z}(t)$ evolves so that $\text{MI}(Z;Y_t)$ increases. Then $\text{MI}(X;Y_t)$ also increases by the data-processing inequality, a classical consequence of Markov refinement [2, 29]:

$$\frac{d}{dt}\text{MI}(X;Y_t) \;\geq\; 0.$$

However, this evolution generally alters the marginal distribution P_{Y_t}, and therefore its entropy $H(Y_t)$ need not remain constant. Consequently,

$$\dot{r}(t) = \dot{H}(X_t) + \dot{H}(Y_t) - 2\,\dot{\mathrm{MI}}(X;Y_t) = \dot{H}(Y_t) - 2\,\dot{\mathrm{MI}}(X;Y_t),$$

since $H(X_t)$ is fixed.

If the change in the marginal entropy is dominated by the MI gain—formally, if there exists $0 \le \alpha < 2$ such that

$$|\dot{H}(Y_t)| \;\le\; \alpha\,\dot{\mathrm{MI}}(X;Y_t),$$

then by the approximate contraction inequality of Section 4,

$$\dot{r}(t) \;\le\; (\alpha - 2)\,\dot{\mathrm{MI}}(X;Y_t) \;<\; 0.$$

Thus, Markovian noise reduction yields contraction of the information radius whenever the marginal drift is sufficiently small relative to the MI increase. Related ideas appear in information-transfer and causality theory, including directed information and transfer entropy [30, 31].

6.4 Gaussian-Mixture Contraction Example

Consider a Gaussian mixture model

$$P_{XY}(x,y) = \tfrac{1}{2}\,\mathcal{N}(x;\mu_1,\Sigma_1)\mathcal{N}(y;\nu_1,\Lambda_1) + \tfrac{1}{2}\,\mathcal{N}(x;\mu_2,\Sigma_2)\mathcal{N}(y;\nu_2,\Lambda_2).$$

Suppose that the mixture parameters evolve so that the components become more aligned in the (X,Y) space—e.g., decreasing separation or reducing component variances. Such alignment typically increases $\mathrm{MI}(X;Y_t)$ [27, 28, 32].

Unlike the Gaussian example with fixed marginals, here the marginal distributions $P_X(t)$ and $P_Y(t)$ generally change over time. Thus

$$\dot{r}(t) = \dot{H}(X_t) + \dot{H}(Y_t) - 2\,\dot{\mathrm{MI}}(X;Y_t),$$

and fixed-marginals contraction does not apply.

If, however, the evolution satisfies the small-drift condition of Section 4—namely, if for some $0 \le \alpha < 2$,

$$|\dot{H}(X_t)| + |\dot{H}(Y_t)| \;\le\; \alpha\,\dot{\mathrm{MI}}(X;Y_t),$$

then the approximate contraction inequality yields

$$\dot{r}(t) \;\le\; (\alpha - 2)\,\dot{\mathrm{MI}}(X;Y_t) \;<\; 0.$$

Geometrically, the mixture components move closer in information space as MI increases, and contraction follows whenever the marginal drift is sufficiently small relative to the MI gain. This provides a nonlinear, multi-modal counterpart to the exact fixed-marginals example and illustrates the role of MI in complex statistical models.

6.5 Applications in Physical, Biological, and Abstract Systems

The contraction of information radius has implications across a wide range of domains in which statistical dependence plays a functional role. We outline several representative application areas.

6.5.1 Physical Systems

In thermodynamic systems, variables often interact through energy-exchange channels that correlate their states. As coupling strengthens, mutual information increases and the uncertainty distance r decreases, consistent with information-transfer principles in nonequilibrium thermodynamics [8, 10, 11, 12].

6.5.2 Biological Networks

In gene regulatory networks, statistical dependence between expression levels provides a signature of coordinated control. Co-regulated genes typically exhibit increasing MI across conditions, reflecting stronger functional linkage [14]. The contraction theorem formalizes this phenomenon: as mutual information grows, the information radius between interacting genes shortens. This view is compatible with copula-based dependence measures in modern systems biology [25].

6.5.3 Distributed Learning and Information Systems

In distributed learning architectures, sensor-fusion frameworks, and recommender systems, shared signal structure leads to increasing MI among components. As common patterns emerge, the system becomes easier to coordinate: components become informationally closer, and global aggregation requires less communication. This perspective connects information geometry with multi-agent learning, echoing analyses of representation learning and SGD dynamics [33, 34, 35].

7 Related Work and Conceptual Connections

The present work sits at the intersection of several established frameworks in information theory, information geometry, and thermodynamic information processing. In this section we summarize the most closely related lines of work and clarify how our contribution differs from and complements them.

Normalized information distances and compression-based metrics. The normalized information distance (NID) and its practical surrogate, the normalized compression distance (NCD), provide universal similarity measures based on Kolmogorov complexity and real-world compression algorithms (e.g., Li et al. [36, 37]). These quantities capture algorithmic similarity between arbitrary objects and have been used widely in clustering, pattern discovery, and anomaly detection.

Our formulation is conceptually related in that it also interprets information as a basis for distance. However, there are two key differences. First, we work entirely within Shannon's probabilistic framework rather than algorithmic complexity [2, 19]. Second, most importantly, NID and NCD are *static* quantities: they measure how far two objects are at a single point in time, but do not describe how the distance evolves under a prescribed dynamics. In contrast, the main theorem of this paper provides a *dynamic* statement: when a system evolves under an MI-gradient flow, the information radius $r = H(X) + H(Y) - 2\,\mathrm{MI}(X;Y)$ contracts monotonically [5, 23]. Thus, our results can be viewed as a thermodynamic and dynamical analogue of information-based distance principles.

Relation to information geometry and statistical distances. In information geometry, the Fisher–Rao metric and KL divergence serve as local and global measures of dissimilarity on manifolds of probability distributions (e.g., Amari and Nagaoka [7], Amari [9]). These constructions endow the space of distributions with a rich differential geometric structure: geodesics, curvature, natural gradients, and projection operators.

Our viewpoint is complementary. We do not propose a new Riemannian metric nor modify classical divergences such as KL. Instead, we fix attention on *pairs* of variables (X, Y) and study the dynamic behavior of the functional

$$r(X, Y) = H(X) + H(Y) - 2\,\mathrm{MI}(X; Y),$$

which can also be written as $H(X|Y) + H(Y|X)$. The contraction theorem shows that under MI-gradient dynamics, r decreases monotonically. While information geometry typically investigates *intrinsic* distances on distribution manifolds, our analysis emphasizes a more *relational* perspective: a dynamic law for how dependence between systems evolves. This view is consistent with geometric treatments of dependence such as copula theory [24, 25], but extends them in a dynamical direction.

Entropy flows and thermodynamic information processing. There is extensive literature on entropy production, information flow, and thermodynamics of computation and feedback control (e.g., Landauer [6], Sagawa and Ueda [10], Horowitz and Esposito [11], Seifert [12], Parrondo et al. [8]). In these studies mutual information often appears as a correction term in generalized second-law inequalities or as a measure of information transfer.

Our formulation is aligned with this tradition but differs in emphasis. We treat MI not only as a measure of dependence or a thermodynamic correction, but as an *active geometric potential* whose gradient generates an informational force field [13]. Viewed in this way, the contraction of the information radius r provides a metric description of how informational coupling shapes the effective geometry of the system.

Summary. To summarize, our contribution can be viewed as a bridge between three perspectives: (i) distance-like measures based on information content, (ii) geometric structures on spaces of probability distributions, and (iii) thermodynamic formulations of information flow. We do not replace these frameworks but instead add a dynamic layer: the MI-gradient induces a contraction of an information radius, providing a geometrically interpretable law for the evolution of dependence between components. In this sense, the present work outlines the dynamical counterpart to classical information geometry and enriches thermodynamic interpretations through the lens of informational dynamics [**?**].

8 Natural Forces as Mutual Dependence

Mutual information (MI) provides a quantitative description of how strongly two systems inform each other. As established in Section 3, whenever the interaction between X and Y evolves along a trajectory that increases $\mathrm{MI}(X_t; Y_t)$, the information radius

$$r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t; Y_t)$$

contracts under the fixed-marginal and small-drift regimes. This section organizes the resulting picture into a unified view of *influential forces* and clarifies how the mechanical analogy extends to general informational geometries [8, 13].

8.1 From Mechanical to Influential Forces

Under the isothermal, closed, quasi-static regime introduced in Section 2, the interaction acquires a mechanical form,

$$\boldsymbol{F}_{att} = k_B T\, \nabla \mathrm{MI}, \qquad \Delta U = -\, k_B T\, \Delta \mathrm{MI}, \tag{8}$$

so an increase in shared information corresponds to a decrease in internal energy. Therefore, within these thermodynamic assumptions, MI-ascent behaves analogously to a classical attractive force [10, 11, 12].

Outside this thermodynamic regime, we use the term *Attractive Influential Force* (AIF) to denote any interaction for which the dynamics satisfy $\dot{\mathrm{MI}}(t) \geq 0$. This broader concept aligns with recent viewpoints in information-thermodynamic coupling and feedback processes [8].

Contraction via MI-ascent. Theorem 3.1 implies that flows aligned with ∇MI necessarily contract $r(t)$ under the two modeling regimes formalized in Section 4. Specifically: (i) in the fixed-marginals regime, $\dot{r}(t) = -2\,\dot{\mathrm{MI}}(t)$, and (ii) in the small-drift regime, $\Delta r \leq (\alpha - 2)\Delta\mathrm{MI} < 0$ when $|\Delta H(X)| + |\Delta H(Y)| \leq \alpha\, \Delta\mathrm{MI}$ for some $0 \leq \alpha < 2$. Thus, the attractive character of AIF does not rely on thermodynamic structure; it follows directly from the informational gradient alone.

Copula consistency (continuous variables). In the continuous case, differential entropies may be negative. To avoid these ambiguities, Section 4.3 adopts a copula-based representation that preserves marginals and expresses

$$\mathrm{MI}(X;Y) = \iint c(u,v) \log c(u,v)\, du\, dv,$$

a classical result in dependence modeling [24, 25, 26]. In this representation, both MI-ascent and the contraction of r carry over verbatim, so the conclusions above remain valid.

8.2 Unified Information Distance and Exponential Kernels

Early formulations of informational influence used the exponential kernel

$$\kappa_r(X,Y) = \exp\!\left(-\tfrac{1}{2}\, r(X,Y)\right), \tag{9}$$

which depends explicitly on $r(X,Y)$.

To accommodate broader settings and model classes, we introduce a unified information radius

$$D_{\mathrm{info}}(X,Y) := \inf_{\Phi,\Psi} D_R\big(\Phi[X],\ \Psi[Y]\big), \tag{10}$$

where D_R denotes an information-geometric divergence on a common reference space and the infimum is taken over admissible representations Φ, Ψ.

Only the structural role of D_{info} is required here; its full development belongs to the broader framework of Influential Force Theory (IFT) [**?**]. Replacing r by D_{info} yields the generalized exponential kernel

$$\kappa_{\mathrm{info}}(X,Y) = \exp\!\left(-\tfrac{1}{2}\, D_{\mathrm{info}}(X,Y)\right), \tag{11}$$

which parallels the role of exponential divergences in information geometry and optimal transport [16].

Remark (continuity and syntactic invariance). The move from κ_r to κ_{info} is not merely cosmetic. It reflects a syntactic invariance of influence structure: under MI-ascent, any distance functional that contracts along the flow and admits an exponential representation induces the same qualitative vector-field alignment with ∇MI. This continuity explains why exponential kernels arise robustly when informational coupling is the driving mechanism.

8.3 Modeling scope, caveats, and identifiability

- **Choice of coordinates.** The gradient ∇MI may refer to finite-dimensional parameters driving $P_{XY,\theta}$ or to distributional directions (Sec. 2). Copula-projected flows (Sec. 4.3) preserve marginals and yield exact contraction identities [25].
- **Non-isothermal and open systems.** In general settings, the mechanical reading $\Delta U = -\,k_B T\,\Delta\text{MI}$ need not hold. Nevertheless, the informational statement (MI-ascent $\Rightarrow$ r-contraction) remains valid under the hypotheses of Sections 3–4, consistent with analyses in stochastic thermodynamics [12].
- **Noise and small-drift regimes.** Bounds in Section 4 quantify robustness: if marginal drift is dominated by MI gain ($|\Delta H(X)| + |\Delta H(Y)| \leq \alpha\,\Delta\text{MI}$ with $0 \leq \alpha < 2$), contraction persists with slack $(\alpha - 2)$. Such robustness parallels results in nonparametric MI estimation and information-flow studies [27, 28].
- **Non-conservative aspects.** Projected gradient flows on constrained manifolds (e.g. the copula manifold) need not be globally integrable to a scalar potential. The contraction statements rely on *directional* MI ascent rather than global potential structure [16].

8.4 Synthesis

The results of Secs. 3–4 establish that MI-ascent contracts r under broad regimes. Naming the generator of this ascent the *attractive influential force* provides a unifying, geometry-first interpretation of "force'' as mutual dependence in action.

The exponential kernels (9)–(11) reveal a deeper syntactic invariance across distance choices and position IFT as a bridge between information geometry, thermodynamic analogies, and dynamical systems, consistent with broader informational coupling principles [**?**].

9 Conceptual Synthesis and Forward Directions

The preceding sections have developed a geometric and thermodynamic interpretation of mutual-information ascent. Section 3 established that when an interaction increases $\text{MI}(X_t;Y_t)$, the information radius

$$r(t) = H(X_t) + H(Y_t) - 2\,\text{MI}(X_t;Y_t)$$

contracts under fixed-marginal or small-drift conditions [5, 23]. Section 4 refined this statement by quantifying robustness to marginal drift and by developing the projected-gradient formulation on the coupling manifold [16, 17]. Section 7 connected these contraction principles to the mechanical and informational notions of the *Attractive Influential Force* (AIF), extending classical ideas in information geometry [7, 9] and thermodynamic feedback processes [10, 11, 12]. The purpose of this section is to synthesize these developments into a cohesive conceptual picture and to describe several avenues for further exploration.

9.1 Interpretive structure of MI-driven contraction

At a high level, MI measures how much the uncertainty of one system is reduced by knowing the other. The distance r captures the portion of uncertainty that remains unexplained. Thus, when $\mathrm{MI}(X_t; Y_t)$ increases along a trajectory, the two systems become more mutually predictive and the unexplained portion shrinks. This contraction serves as the geometric signature of increasing informational alignment, a phenomenon consistent with dependence-based geometries such as copula representations [24].

This alignment is driven by the local behavior of $\nabla\mathrm{MI}$. Theorem 3.1 shows that if the interaction follows this gradient under the fixed-marginal or small-drift regimes, then $r(t)$ decreases monotonically. This places MI at the center of an information-geometric mechanism that organizes the evolution of interacting systems [7, 9].

9.2 Unified exponential structures and informational geometry

Section 7 introduced a unified information distance D_{info}, defined in structural form in (10). Early formulations employed the exponential kernel $\kappa_r = \exp(-\frac{1}{2}r)$, while the generalized form $\kappa_{\mathrm{info}} = \exp(-\frac{1}{2}D_{\mathrm{info}})$ preserves the same qualitative alignment with $\nabla\mathrm{MI}$. This reflects a syntactic invariance across admissible distance choices and echoes the role of exponential divergences in broader information-geometric and optimal-transport contexts [16].

9.3 Modeling scope, caveats, and identifiability

The contraction principles rely on assumptions that warrant careful interpretation.

- **Choice of coordinates.** The gradient $\nabla\mathrm{MI}$ may refer to finite-dimensional parameters or to distributional directions (Sec. 2). Copula-projected flows preserve marginals and yield exact contraction identities [25].
- **Non-isothermal and open systems.** In general thermodynamic settings, the mechanical correspondence $\Delta U = -k_B T\,\Delta\mathrm{MI}$ need not hold. Nevertheless, the informational statement (MI-ascent $\Rightarrow$ r-contraction) remains valid under the hypotheses of Sections 3-4 [10, 11].
- **Noise and small-drift regimes.** Bounds in Section 4 quantify robustness: when marginal drift is dominated by MI gain, contraction persists with slack $(\alpha - 2)$. This perspective is compatible with modern statistical approaches to MI estimation and smoothing [27, 28].
- **Directional and non-conservative aspects.** Projected gradient flows need not be globally integrable to a scalar potential; they rely on *directional* MI ascent rather than global potential structure. Related ideas appear in directional-information and causal dependence measures, such as directed information and transfer entropy [29, 30].

9.4 Application arenas and tests

Several empirical settings offer opportunities for evaluating the contraction principle. In learning dynamics, tracking $r(t)$ and $\mathrm{MI}(X_t; Y_t)$ during representation training may reveal phases of increasing informational alignment. This perspective provides conceptual bridges to analyses of SGD dynamics and representation learning [33, 34, 35].

Biological and neural systems provide additional testbeds where coupling strength or noise levels can be experimentally manipulated. Increasing the reliability of a stimulus channel or strengthening synaptic coupling, for example, is expected to increase MI and reduce $r(t)$, consistent with the contraction theorem. Such experiments relate naturally to information-flow analyses in systems neuroscience [14].

In multi-agent or networked systems, monitoring pairwise and multi-way MI during coordinated interaction episodes may reveal network-level contraction effects. Directional or asymmetric interactions (e.g., transfer-entropy-based) may further modulate the rate and extent of contraction [30, 31].

A further empirical direction concerns gene-regulatory systems. Earlier work introduced the *ab initio Genetic Orbital (GO) method* and derived information-geometric embeddings of gene networks. Since MI-ascent is expected to shorten information radii within such clusters, the contraction principle provides a theoretical basis for interpreting these structures as stable attractors. Extensions such as the Ab Initio Network Expander (AINE) may enable systematic validation of informational forces in biological data [**?**].

9.5 Conclusion

Taken together, these perspectives outline a program in which MI-ascent provides the geometric driver of coupling, exponential structures furnish a unifying kernel layer, and D_{info} extends this picture to broader informational geometries. These themes form a conceptual bridge to the concluding section and point toward future directions in the IFT program [**?**].

10 Summary

This work has developed a geometric and thermodynamic interpretation of mutual information (MI) ascent between interacting systems. The central object is the information radius

$$r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(X_t; Y_t),$$

which measures uncertainty not shared between the systems and can be written as $H(X_t|Y_t) + H(Y_t|X_t)$ in the discrete case [2]. When the interaction induces an increase in $\mathrm{MI}(X_t; Y_t)$, the distance $r(t)$ contracts, connecting MI gradients to attractive informational forces and providing a framework in which information-driven interactions can be studied geometrically.

10.1 Summary of core findings

Section 3 established a contraction theorem for the distance $r(t)$ under MI-ascent dynamics [5, 23]. Assuming an underlying evolution (X_t, Y_t) for which $\mathrm{MI}(X_t; Y_t)$ is differentiable in time and the dynamics are governed by ascent along $\nabla\mathrm{MI}$, Theorem 3.1 shows that $\dot{r}(t) \leq 0$. Section 4 refined this statement by isolating two principal regimes: (i) the fixed-marginal regime, yielding the exact identity $\dot{r}(t) = -2\,\dot{\mathrm{MI}}(t)$, and (ii) the small-drift regime, where contraction persists whenever marginal drift is dominated by the MI increase.

Section 5 interpreted these results in an information-geometric language, viewing MI ascent as inducing a contractive vector field on an abstract space of informational states [7, 9]. Section 4.3 extended the analysis to continuous variables via a copula-based representation that preserves marginals and expresses $\mathrm{MI}(X; Y)$ as an integral of $c(u, v) \log c(u, v)$ on $[0, 1]^2$.

Section 7 introduced the *Attractive Influential Force* (AIF) as the conceptual mechanism behind these results. In its mechanical form, valid under isothermal, closed, quasi-static conditions, the interaction obeys

$$\boldsymbol{F}_{att} = k_B T\,\nabla\mathrm{MI}, \qquad \Delta U = -\,k_B T\,\Delta\mathrm{MI},$$

reflecting classical correspondences in information thermodynamics [10, 11, 12]. In its general informational form, AIF refers to any interaction whose flow satisfies $\dot{\mathrm{MI}}(t) \geq 0$, a viewpoint that extends naturally to the variational formulation on the coupling space [16, 17].

10.2 Conceptual implications

Conceptually, the results of this paper show that MI plays a dual role. As a scalar measure, it quantifies statistical dependence. As a geometric functional, it defines a potential whose gradient generates contractive flows. The distance $r(t)$ corresponds to the portion of uncertainty not explained by the interaction, and its contraction describes increasing informational alignment.

The thermodynamic analogy complements this geometric view. Under isothermal, closed, quasi-static conditions, the relationship $\Delta U = -k_B T\, \Delta\mathrm{MI}$ offers an energetic interpretation of MI ascent. However, this analogy is not essential: the contraction theorem persists under broader assumptions that do not rely on thermodynamic structure.

Exponential kernels such as $\kappa_r = \exp(-\frac{1}{2}r)$ and its generalized form $\kappa_{\mathrm{info}} = \exp(-\frac{1}{2}D_{\mathrm{info}})$ further reveal a syntactic invariance across distance choices. This suggests robustness of informational influence structures beyond the specific form of r, and connects to broader information-geometric and optimal-transport perspectives [16].

10.3 Placement within Influential Force Theory

Although the present paper can be read independently, it also forms part of a broader program referred to as *Influential Force Theory* (IFT), which interprets natural forces as manifestations of informational interactions. Within this program, the distance $r(t)$ represents a tractable discrete-time instance of a more general unified distance D_{info}. The definitions and kernels introduced in Section 7 illustrate how MI gradients, distance contraction, and exponential forms arise as facets of a single structural motif [**?**].

In this perspective, the Attractive Influential Force (AIF) is not a stand-alone construct but a particular realization of the more general principle that MI-ascent induces geometric contraction. The unified distance D_{info} extends this idea to situations in which r is not directly applicable or where richer representational hierarchies are required.

10.4 Applications and empirical frontiers

Applications span learning dynamics, biological information flow, network coordination, and gene-regulatory modeling. In learning systems, monitoring $r(t)$ and $\mathrm{MI}(X_t; Y_t)$ during training may reveal phases of increasing informational alignment, echoing recent analyses of representation learning and SGD dynamics [33, 34]. Similarly, biological and neural systems provide testbeds in which changes in coupling strength or noise levels produce measurable MI increases [14]. Networked and multi-agent systems offer another setting in which pairwise or multi-way MI may contract during coordinated behavior.

A further empirical direction concerns gene-regulatory networks, where MI-based geometric embeddings (e.g., Genetic Orbital and Network Expander methods) suggest stable informational attractors whose structure is consistent with the contraction theorem [**?**].

10.5 Outlook

Several directions for future work emerge from this analysis. A theoretical question concerns the characterization of the class of distance functionals for which MI ascent guarantees contraction and for which exponential kernels remain aligned with $\nabla\mathrm{MI}$. Another avenue is the extension of the framework to repulsive regimes and externally driven systems, where non-informational fluxes may dominate the net evolution.

A further direction concerns causality and directionality. Many systems exhibit inherently directed influence, and an open question is whether analogous contraction results exist for directional distances constructed from causal measures such as directed information or transfer entropy.

In summary, MI-ascent defines a geometric mechanism of informational contraction and supports a framework that unifies mechanical analogies, distance-based representations, and exponential influence structures. These results provide a foundation on which a broader theory of informational forces may be built, and the IFT program seeks to develop that theory systematically [?].

Appendix A: Rigorous Formulation of the MI-Gradient Flow

This appendix provides a rigorous variational formulation of the mutual-information gradient flow and the corresponding contraction theorem. The aim is to place the main result of Section 3 on a precise geometric and analytic foundation, consistent with classical treatments of gradient flows in spaces of probability measures [16, 17, 18].

A.1 Coupling Space and Fixed Marginals

Let $\mu \in \Delta^n$ and $\nu \in \Delta^m$ be fixed marginal distributions associated with random variables X and Y. Define the coupling space

$$\Pi(\mu, \nu) = \left\{ P \in \mathbb{R}_+^{n \times m} : P\mathbf{1} = \mu,\ P^\top \mathbf{1} = \nu \right\}.$$

For any $P \in \Pi(\mu, \nu)$, the mutual information is

$$\mathrm{MI}(P) = D_{\mathrm{KL}}\Big(P \,\|\, \mu\nu^\top\Big),$$

the classical KL-based representation of MI. The associated information radius is

$$r(P) = H(\mu) + H(\nu) - 2\,\mathrm{MI}(P),$$

which coincides with $H(X|Y) + H(Y|X)$ in the discrete case.

A.2 Projected Gradient Flow with Frobenius Geometry

The ambient space $\mathbb{R}^{n \times m}$ is equipped with the Frobenius inner product

$$\langle A, B \rangle = \sum_{i,j} a_{ij} b_{ij}.$$

At any $P \in \Pi(\mu, \nu)$, the tangent space is

$$T_P = \Big\{ \Delta \in \mathbb{R}^{n \times m} : \Delta\mathbf{1} = 0,\ \Delta^\top \mathbf{1} = 0 \Big\}.$$

Assumption 3 (Projected MI-gradient flow on $\Pi(\mu, \nu)$)**.** *The dynamics of the joint distribution satisfy*

$$\dot{P}(t) = \eta\, \Pi_{T_{P(t)}}\big(\nabla \mathrm{MI}(P(t))\big), \qquad \eta > 0,$$

where $\Pi_{T_{P(t)}}$ denotes the Frobenius-orthogonal projection onto $T_{P(t)}$. This construction parallels the variational structure of gradient flows in optimal-transport geometry [16, 17, 18].

A.3 Gâteaux Derivative and Gradient of MI

Lemma A.1 (Gâteaux derivative of MI)**.** *For any feasible direction $\Delta \in T_P$,*

$$\mathrm{DMI}[P](\Delta) = \Big\langle \log \frac{P}{\mu\nu^\top} + \mathbf{1},\ \Delta \Big\rangle,$$

where the logarithm and division act elementwise and $\mathbf{1}$ denotes the all-ones matrix. Consequently,

$$\nabla \mathrm{MI}(P) = \log \frac{P}{\mu\nu^\top} + \mathbf{1},$$

consistent with classical variational expressions for KL-type divergences.

A.4 Contraction of the Information Radius

Theorem A.2 (Contraction of information radius)**.** *Under Assumption 3, the information radius $r(P(t))$ is monotonically non-increasing:*

$$\frac{d}{dt}\, r(P(t)) \le 0.$$

Moreover, equality holds at time t if and only if

$$\Pi_{T_{P(t)}}\big(\nabla \mathrm{MI}(P(t))\big) = 0.$$

Proof. Since the marginals μ and ν remain fixed, $H(\mu)$ and $H(\nu)$ are constant in t. Therefore,

$$\frac{d}{dt} r(P(t)) = -2\, \frac{d}{dt} \mathrm{MI}(P(t)).$$

By the chain rule and the gradient-flow dynamics,

$$\begin{aligned}
\frac{d}{dt}\mathrm{MI}(P(t)) &= \Big\langle \nabla \mathrm{MI}(P(t)),\, \dot{P}(t) \Big\rangle \\
&= \eta \left\langle \nabla \mathrm{MI}(P(t)),\, \Pi_{T_{P(t)}}(\nabla \mathrm{MI}(P(t))) \right\rangle \\
&= \eta \left\| \Pi_{T_{P(t)}}(\nabla \mathrm{MI}(P(t))) \right\|_F^2 \;\ge\; 0.
\end{aligned}$$

Thus $r(P(t))$ is non-increasing, with equality precisely when the projected gradient vanishes. □

A.5 Remarks on Continuous Analogues

Remark A.3. (1) The coupling-space formulation avoids ambiguities arising from taking gradients with respect to state values of discrete random variables.

(2) A continuous analogue can be formulated on the manifold of joint densities with fixed marginals, using either the L^2 metric or the Fisher–Rao geometry. The resulting projected-gradient formulation again yields monotonicity of MI and the associated contraction of r.

(3) Copula-based parameterizations provide an alternative dependence-only representation of MI, consistent with the continuous formulations described in Section 4.

Appendix B: Detailed Proofs

This appendix collects supplementary proofs of the results stated in Section 4. The arguments rely on standard properties of Shannon entropy, mutual information, and copula-based dependence representations [2, 24, 25].

B.1 Proof of Proposition 4.3 (Discrete-Step Contraction)

Assumptions. Let $(X, Y) \mapsto (X', Y')$ be a one-step transition with finite entropies. Assume $\Delta\mathrm{MI} > 0$ and that there exists $\alpha \in [0, 2)$ such that

$$|\Delta H(X)| + |\Delta H(Y)| \leq \alpha\, \Delta\mathrm{MI}.$$

Step 1: Exact expansion of Δr. With $r(X, Y) := H(X) + H(Y) - 2\,\mathrm{MI}(X; Y)$,

$$\Delta r = r(X', Y') - r(X, Y) = \Delta H(X) + \Delta H(Y) - 2\,\Delta\mathrm{MI}.$$

Step 2: Dominance bound. Using the dominance condition,

$$\Delta r \leq (|\Delta H(X)| + |\Delta H(Y)|) - 2\,\Delta\mathrm{MI} \leq (\alpha - 2)\,\Delta\mathrm{MI} < 0.$$

This proves contraction in the discrete step.

Remark (Tightness). If $\Delta H(X) = \Delta H(Y) = 0$, then $\Delta r = -2\Delta\mathrm{MI}$, achieving the ideal reduction in the copula-only case. Such inequalities parallel the bounds used in MI stability and estimation theory [27, 28]. □

B.2 Proof of Proposition 4.5 (Noise-Robust Differential Form)

Assumptions. Assume $t \mapsto P_{XY,t}$ is differentiable and entropies are finite along the path. Suppose a measurable function $\beta(t) \geq 0$ satisfies

$$|\dot{H}(X_t)| + |\dot{H}(Y_t)| \leq \beta(t).$$

Step 1: Differential identity. From $r(t) = H(X_t) + H(Y_t) - 2\,\mathrm{MI}(t)$,

$$\dot{r}(t) = \dot{H}(X_t) + \dot{H}(Y_t) - 2\,\dot{\mathrm{MI}}(t) \leq \beta(t) - 2\,\dot{\mathrm{MI}}(t).$$

Step 2: Quantitative contraction. If $\dot{\mathrm{MI}}(t) \geq \frac{\beta(t)}{2} + \gamma$ on an interval $I = [t_1, t_2]$ for some $\gamma > 0$, then

$$\dot{r}(t) \leq -2\gamma, \qquad t \in I,$$

and integrating gives

$$r(t_2) - r(t_1) \leq -2\gamma\,(t_2 - t_1).$$

Thus r decreases at rate at least γ despite small marginal drifts. Such robustness parallels stability analyses in MI-based estimators and information-flow studies [27, 28]. □

B.3 Copula-Only Flow Identities and Monotonicity

Setup and regularity. Let $c_s \in \mathcal{C}$ be a C^1 path of copula densities on $[0,1]^2$ with $c_s(u,v) \geq \varepsilon > 0$, $\iint c_s \, du \, dv = 1$, and $\int c_s(u,v)\, dv = \int c_s(u,v)\, du = 1$ for all s. Write

$$p_s(x,y) = f_X(x) f_Y(y) c_s(F_X(x), F_Y(y)),$$

with fixed marginals f_X, f_Y.

Step 1: Differentiation under the integral. Since $\mathrm{MI}(c) = \iint c \log c$, for Gâteaux variations δc satisfying the copula constraints [24, 25],

$$\delta \, \mathrm{MI}[c] = \iint \delta c \, (1 + \log c).$$

Hence,

$$\frac{d}{ds}\mathrm{MI}_s = \iint \dot{c}_s(u,v) \, (1 + \log c_s(u,v)) \, du \, dv.$$

Step 2: Tangent-space constraints. Marginal preservation implies

$$\int_0^1 \dot{c}_s(u,v) \, du = 0, \qquad \int_0^1 \dot{c}_s(u,v) \, dv = 0,$$

so $\dot{c}_s \in T_{c_s}\mathcal{C}$, the copula tangent space described in classical treatments of dependence geometry [26, 25].

Step 3: Projected gradient flow and positivity. Consider the constrained L^2-gradient flow

$$\dot{c}_s = \Pi_{T_{c_s}}\big(1 + \log c_s\big).$$

Then

$$\frac{d}{ds}\mathrm{MI}_s = \big\langle \Pi_{T_{c_s}}(1 + \log c_s), \, (1 + \log c_s) \big\rangle = \big\| \Pi_{T_{c_s}}(1 + \log c_s) \big\|_2^2 \; \geq \; 0,$$

showing monotonic increase of MI along the projected gradient flow. This argument parallels the variational structure appearing in information–geometric and optimal-transport analyses [16].

Step 4: Consequence for r. Because $H(X)$ and $H(Y)$ are fixed under copula flows,

$$\frac{d}{ds} r(s) = -2 \, \frac{d}{ds}\mathrm{MI}(s) \leq 0,$$

recovering the exact contraction property under fixed marginals. □

Appendix C: Natural Forces and Information-Theoretic Structure

In this appendix we summarize a class of dynamics referred to as natural forces and record two structural results. These results connect classical notions of mean force and marginal preservation with the informational formulation developed in the main text. The viewpoint aligns with modern interpretations of information-thermodynamic feedback and nonequilibrium processes.

C.1 Preliminaries

Let

$$X = (U, Q, P), \qquad Y = (V, R, S),$$

where U and V are intrinsic variables and (Q, P), (R, S) are extrinsic variables. Let $p_t(x, y)$ denote the joint density at time t and define the time-dependent mutual information

$$\mathrm{MI}_t(X;Y) := \int p_t(x, y) \log\left(\frac{p_t(x, y)}{p_t(x)\, p_t(y)}\right) dx\, dy.$$

Define the intrinsic entropies

$$H(X) := H(U), \qquad H(Y) := H(V),$$

and the associated information radius

$$r_{\mathrm{info}}(t) := H(X) + H(Y) - 2\,\mathrm{MI}_t(X;Y).$$

Definition (Natural Force Class $\mathcal{N}$). Let L denote the generator of the dynamics and L^* its adjoint. We say the dynamics belong to the natural force class $\mathcal{N}$ if L admits a decomposition

$$L = L_U^{\mathrm{int}} + L_V^{\mathrm{int}} + L_{Q,P}^{\mathrm{ext}} + L_{R,S}^{\mathrm{ext}} + K_{\mathrm{rel}},$$

where:

(i) K_{rel} is induced by a central potential $V(r)$ with $r = Q - R$;

(ii) the relational term preserves intrinsic marginals:

$$\Pi_U \circ K_{\mathrm{rel}}^* \equiv 0, \qquad \Pi_V \circ K_{\mathrm{rel}}^* \equiv 0;$$

(iii) the initial intrinsic marginals $p_0(u)$ and $p_0(v)$ are invariant under L_U^{int} and L_V^{int}.

These conditions match classical formulations of Fokker–Planck generators.

C.2 Preservation of Intrinsic Marginal Entropy

Theorem C.1 (Natural forces preserve intrinsic marginal entropy)**.** *Under the natural force class $\mathcal{N}$, for all $t \geq 0$,*

$$p_t(u) \equiv p_0(u), \qquad p_t(v) \equiv p_0(v),$$

and hence

$$H(U_t) = H(U_0), \qquad H(V_t) = H(V_0).$$

Sketch of proof. From the definition,

$$\partial_t p_t(u) = \Pi_U\big(L^* p_t\big).$$

Because $K_{\rm rel}^*$ annihilates intrinsic marginals and $p_0(u)$ is invariant under $L_U^{\rm int}$, it follows that $\partial_t p_t(u) = 0$. A symmetric argument applies to v.

C.3 Potential of Mean Force and Pointwise MI

Let r denote a relative coordinate (e.g., separation). Define the pointwise mutual information

$$\mathrm{pmi}(r) := \log\left(\frac{p(r)}{p_X(r)\, p_Y(r)}\right).$$

Theorem C.2 (Potential of mean force equals MI gradient)**.** *In the canonical ensemble at temperature T, under the natural-force class and stationary one-body marginals,*

$$F_{\rm mean}(r) = k_B T\, \nabla_r \mathrm{pmi}(r).$$

Sketch of proof. At equilibrium,

$$g(r) = \frac{p(r)}{p_X(r) p_Y(r)} = e^{-\beta W(r)}, \qquad \beta = \frac{1}{k_B T}.$$

Taking logarithms yields $\mathrm{pmi}(r) = -\beta W(r)$, hence

$$-\nabla_r W(r) = k_B T\, \nabla_r \mathrm{pmi}(r),$$

the classical mean-force identity.

C.4 Information-Distance Contraction under Natural Forces

Assume the intrinsic entropies remain constant as guaranteed by Theorem C.1. If

$$\frac{d}{dt}\mathrm{MI}_t(X;Y) \geq 0,$$

then

$$\frac{d}{dt} r_{\rm info}(t) = -2\,\frac{d}{dt}\mathrm{MI}_t(X;Y) \;\leq\; 0.$$

This parallels the fixed-marginals case in Proposition 4.10, and shows that MI-increasing dynamics within $\mathcal{N}$ necessarily contract the information radius.

C.5 Non-Equilibrium Corrections

Remark. General nonequilibrium systems allow a decomposition of the probability current

$$J = J_{\rm grad} + J_{\rm rot},$$

where only $J_{\rm grad}$ aligns with $\nabla\mathrm{MI}$. This yields refined identities of the form

$$\frac{d}{dt} r_{\rm info}(t) = -2\,\frac{d}{dt}\mathrm{MI}_t(X;Y) + O(\|J_{\rm rot}\|)\,,$$

consistent with nonequilibrium thermodynamic formulations of entropy production and information flow. The limit $J_{\rm rot} \to 0$ recovers perfect marginal-entropy preservation.

Appendix D: Full IFT Syntactic Items

This appendix collects the formal versions of several syntactic items used throughout the paper. These structures belong to the broader Influential Force Theory (IFT) framework [5, ?] and serve as information-geometric counterparts to energetic and probabilistic interaction principles.

Units. All entropies and mutual information are measured in nats.

Index of Syntactic Items

The following subsections present complete statements.

D.1 Theorem D.1: Distance Contraction under MI-Ascent

Let (X_t, Y_t) be jointly evolving random variables with joint density $p_t(x, y)$. Define

$$\mathrm{MI}_t := \mathrm{MI}(X_t; Y_t) = \int p_t(x, y) \log \frac{p_t(x, y)}{p_t(x)p_t(y)} \, dx \, dy,$$

and the information radius

$$r(t) := H(X_t) + H(Y_t) - 2\,\mathrm{MI}_t.$$

Assume:

(i) (X_t, Y_t) evolve differentiably in t;

(ii) the flow satisfies $\dot{\mathrm{MI}}_t \geq 0$;

(iii) either:

- *Fixed marginals:* $H(X_t)$ and $H(Y_t)$ are constant, or
- *Small-drift regime:* there exists $0 \leq \alpha < 2$ such that

$$|\Delta H(X)| + |\Delta H(Y)| \leq \alpha \, \Delta\mathrm{MI}.$$

Conclusion.

- Under fixed marginals,

$$\dot{r}(t) = -2\,\dot{\mathrm{MI}}_t \leq 0.$$

- Under small marginal drift,

$$\Delta r \leq (\alpha - 2)\,\Delta\mathrm{MI} < 0.$$

This theorem formalizes the contraction phenomenon, consistent with entropy–divergence inequalities [23] and the information-geometric interpretation of dependence [7, 9].

D.2 Remark D.2: Approximate Distance Contraction

Let

$$E = \Delta r + 2\,\Delta\mathrm{MI} = \Delta H(X) + \Delta H(Y).$$

If for $0 \le \varepsilon < 2$,

$$|\Delta H(X)| + |\Delta H(Y)| \le \varepsilon\,|\Delta\mathrm{MI}|,$$

then

$$\frac{|E|}{2|\Delta\mathrm{MI}|} \le \frac{\varepsilon}{2}.$$

Replacing Δr by $-2\Delta\mathrm{MI}$ thus incurs a relative error bounded by $\varepsilon/2$, consistent with stability analyses of MI estimators and divergence functionals [2].

D.3 Lemma D.3: Marginal-Preserving Generators

Consider a smooth path $s \mapsto p_s(x, y)$ with associated copula densities $c_s(u, v)$ [24, 25]. Assume the marginal constraints

$$\int \dot{p}_s(x, y)\,dy = 0, \qquad \int \dot{p}_s(x, y)\,dx = 0$$

are preserved, and that $c_s(u, v) > 0$ and sufficiently regular.

Then

$$\dot{p}_s(x, y) = f_X(x)\,f_Y(y)\,\dot{c}_s(u, v),$$

where $\dot{c}_s$ satisfies the copula tangent constraints

$$\int_0^1 \dot{c}_s(u, v)\,du = 0, \qquad \int_0^1 \dot{c}_s(u, v)\,dv = 0.$$

Furthermore,

$$\frac{d}{ds}\mathrm{MI}_s = \iint \dot{c}_s(u, v)\,\log c_s(u, v)\,du\,dv,$$

consistent with KL-type variational identities and dependence-specific representations [26].

D.4 Theorem D.4: Energy Flow via Mutual Information

In the canonical ensemble at temperature T, let r be a relative variable with pair distribution

$$g(r) := \frac{p(r)}{p_X(r)p_Y(r)} = e^{-\beta W(r)}, \qquad \beta = \frac{1}{k_B T}.$$

Then the mean force is

$$\mathbf{F}_{\mathrm{mean}}(r) = k_B T\,\nabla_r \log g(r) = k_B T\,\nabla_r \mathrm{pmi}(r),$$

where $\mathrm{pmi}(r) = \log g(r)$.

This identity is classical in statistical mechanics and mean-force theory, appearing in standard references [40, 41]. Its expression as an MI gradient reflects energetic interpretations in information thermodynamics [10, 11, 13].

D.5 Proposition D.5: Force as Energy per Unit MI

Define the informational potential

$$U_{\text{info}} := -\, k_B T \, \text{MI}.$$

Then the associated informational force is

$$\boldsymbol{F}_{att} = -\nabla U_{\text{info}} = k_B T \, \nabla \text{MI}.$$

More generally, for any process satisfying

$$\Delta U = -\, k_B T \, \Delta \text{MI},$$

the force aligns with ∇MI, representing the energetic cost (or gain) per unit increase in mutual information.

This structural analogy between energetic and informational gradients underlies the Attractive Influential Force in IFT and connects classical energy-based formulations with informational coupling mechanisms [8, **?**].

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948. https://ieeexplore.ieee.org/abstract/document/6773024

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, 2006. https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X

[3] R. M. Gray, *Entropy and Information Theory*, Springer, 2011. https://link.springer.com/book/10.1007/978-1-4419-7970-4

[4] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial," *Found. Trends Commun. Inf. Theory*, 1, 2004. https://www.nowpublishers.com/article/Details/CIT-004

[5] T. Mori, "Mutual Information Gradient Induces Contraction of Information Radius," Jxiv preprint, submitted 2025.

[6] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. Develop.*, 5(3), 1961. https://doi.org/10.1147/rd.53.0183

[7] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Amer. Math. Soc., 2000. https://www.amazon.com/Methods-Information-Geometry-Monographs-Statistics/dp/0821843028

[8] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, "Thermodynamics of Information," *Nature Physics*, 11, 2015. https://doi.org/10.1038/nphys3230

[9] S. Amari, *Information Geometry and Its Applications*, Springer, 2016. https://link.springer.com/book/10.1007/978-4-431-55978-8

[10] T. Sagawa and M. Ueda, "Second Law of Thermodynamics with Discrete Quantum Feedback Control," *Phys. Rev. Lett.*, 100:080403, 2008. https://doi.org/10.1103/PhysRevLett.100.080403

[11] J. M. Horowitz and M. Esposito, "Thermodynamics with Continuous Information Flow," *Phys. Rev. X*, 4:031015, 2014. https://doi.org/10.1103/PhysRevX.4.031015

[12] U. Seifert, "Stochastic thermodynamics, fluctuation theorems and molecular machines," *Rep. Prog. Phys.*, 75, 126001, 2012. https://doi.org/10.1088/0034-4885/75/12/126001

[13] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, "Thermodynamics of Prediction," *Phys. Rev. Lett.*, 109:120604, 2012. https://doi.org/10.1103/PhysRevLett.109.120604

[14] M. Deweese and W. Bialek, "Information Flow in Sensory Neurons," *Neural Computation*, 18, 2006. https://link.springer.com/article/10.1007/BF02451830

[15] T. Mori, "Influential Force: From Higgs to the Novel Immune Checkpoint KYNU," Jxiv preprint, 2022. https://doi.org/10.51094/jxiv.156

[16] C. Villani, *Optimal Transport: Old and New*, Springer, 2008. https://doi.org/10.1007/978-3-540-71050-9

[17] F. Otto, “The Geometry of Dissipative Evolution Equations,” *Commun. PDE*, 26(1-2), 2001. https://doi.org/10.1081/PDE-100002243

[18] R. Jordan, D. Kinderlehrer, and F. Otto, “Variational Formulation of the Fokker–Planck Equation,” *SIAM J. Math. Anal.*, 29(1), 1998. https://doi.org/10.1137/S0036141096303359

[19] S. Kullback, *Information Theory and Statistics*, Dover, 1959. https://store.doverpublications.com/products/9780486696843

[20] S. Amari, “Natural Gradient Works Efficiently in Learning,” *Neural Computation*, 10(2), 1998. https://doi.org/10.1162/089976698300017746

[21] N. Tishby, F. Pereira, and W. Bialek, “The Information Bottleneck Method,” Proc. Allerton Conf., 1999. https://www.cs.huji.ac.il/labs/learning/Papers/allerton.pdf

[22] R. Shwartz-Ziv and N. Tishby, “Opening the Black Box of Deep Neural Networks via Information,” arXiv:1703.00810, 2017. https://arxiv.org/abs/1703.00810

[23] I. Csiszár, “I-Divergence Geometry and Minimization Problems,” *Ann. Probab.*, 3(1), 1975. https://www.jstor.org/stable/2959270

[24] A. Sklar, “Fonctions de Répartition à n Dimensions,” Publ. Inst. Statist. Univ. Paris, 1959. https://hal.science/hal-04094463/document

[25] R. Nelsen, *An Introduction to Copulas*, 2nd ed., Springer, 2006. https://link.springer.com/book/10.1007/0-387-28678-0

[26] H. Joe, *Multivariate Models and Dependence Concepts*, CRC Press, 1997. https://doi.org/10.1201/9780367803896

[27] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating Mutual Information,” *Phys. Rev. E*, 69:066138, 2004. https://doi.org/10.1103/PhysRevE.69.066138

[28] M. Belghazi et al., “Mutual Information Neural Estimation,” *ICML*, 2018. https://proceedings.mlr.press/v80/belghazi18a.html

[29] J. Massey, “Causality, Feedback and Directed Information,” Proc. ISIT, 1990. https://isiweb.ee.ethz.ch/papers/arch/mass-1990-1.pdf

[30] T. Schreiber, “Measuring Information Transfer,” *Phys. Rev. Lett.*, 85:461, 2000. https://doi.org/10.1103/PhysRevLett.85.461

[31] L. Barnett, A. K. Seth, and A. B. Barrett, “Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables,” *Phys. Rev. Lett.*, 103:238701, 2009. https://doi.org/10.1103/PhysRevLett.103.238701

[32] L. Kozachenko and N. Leonenko, “Sample Estimate of the Entropy of a Random Vector,” *Probl. Inf. Transm.*, 23, 1987. https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=ppi&paperid=797&option_lang=eng

[33] A. Saxe, J. McClelland, and S. Ganguli, “Exact Solutions to the Nonlinear Dynamics of Deep Linear Neural Networks,” arXiv:1312.6120, 2013. https://arxiv.org/abs/1312.6120

[34] S. Mei, A. Montanari, and P.-M. Nguyen, "Dynamics of Stochastic Gradient Descent for Two-Layer Neural Networks in the Teacher-Student Setup," *NeurIPS*, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/hash/cab070d53bd0d200746fb852a922064a-Abstract.html

[35] A. Achille and S. Soatto, "Information Dropout: Learning Optimal Representations via Information Bottleneck," *IEEE Trans. PAMI*, 2018. https://doi.org/10.1109/TPAMI.2017.2784440

[36] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi, "The Similarity Metric," *IEEE Trans. Inf. Theory*, 50(12), 2004. https://doi.org/10.1109/TIT.2004.838101

[37] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 4th ed., Springer, 2019. https://link.springer.com/book/10.1007/978-3-030-11298-1

[38] C. W. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th ed., Springer, 2009. https://link.springer.com/book/9783540707127

[39] H. Risken, *The Fokker–Planck Equation: Methods of Solution and Applications*, 2nd ed., Springer, 1989. https://link.springer.com/book/10.1007/978-3-642-61544-3

[40] D. A. McQuarrie, *Statistical Mechanics*, Univ. Science Books, 2000. https://mitpress.mit.edu/9781891389153/statistical-mechanics

[41] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford Univ. Press, 2010. https://global.oup.com/academic/product/statistical-mechanics-theory-and-molecular-simulation-9780198825562