Distinction Geometry: Projection, Reconstruction Defect, and the Interface-Native Metric

Mathematical Core

Flyxion

Independent Researcher

June 23, 2026

Abstract

We develop the mathematical core of a framework in which the primitive object of inquiry is not a hidden state space but the geometry of distinguishability between observable projections. Three interlocking structures are established. First, the *reconstruction defect* $\Delta_R(h) = d(h, R(\Pi(h)))$ is introduced as a unified measure of irreversibility: the persistent failure of any map from observations back to the generating history. We prove that irreversibility, in this sense, is equivalent to the non-injectivity of the projection Π, and that the minimum achievable reconstruction defect is controlled by the information lost under projection. Second, the *distinction functional* $\mathcal{D}(P, Q) = \mathrm{JSD}(P\|Q)$ is developed as the canonical measure of the information destroyed when two observable interfaces are merged. Its key properties are established: interface nativity, the entropy decomposition, the gauge-orbit characterisation, and the operationally precise meaning of the two entropy terms. Third, the *space of observable interfaces* $(\mathcal{M}, d_{\mathrm{int}})$ is shown to be a metric space that is locally Riemannian with Fisher information metric. Its curvature is interpreted: positive curvature corresponds to regions where many internal models produce nearly identical observations (overfitting, paradigm stability, institutional opacity); negative curvature to regions of rapid observational divergence (high falsifiability). These three structures — reconstruction defect, distinction functional, and interface geometry — form a mathematical core from which physical, cognitive, and institutional applications follow as corollaries.

Contents

Orientation **3**

1 The Projection Setting **3**
1.1 Histories, Projections, and Fibers 3
1.2 Admissibility and the Projection Sufficiency Principle 4

2 Reconstruction Defect and Irreversibility **5**
2.1 The Reconstruction Problem 5
2.2 The Defect Spectrum 6
2.3 Residual Uncertainty and Reconstruction Defect 7
2.4 Irreversibility as Accumulated Defect 8

3 The Distinction Functional **9**
3.1 The Problem of Comparing Interfaces 9
3.2 Definition and Basic Properties 9
3.3 The Distinction Functional 10
3.4 Interface-Nativity 11
3.5 Gauge Orbits and Hidden Directions 12

4 The Geometry of Interface Space **13**
4.1 The Space of Theories as a Metric Space 13
4.2 Local Riemannian Structure: The Fisher Metric 14
4.3 Curvature and Its Interpretation 15
4.4 Geodesics as Paths of Minimal Accumulated Distinction 16
4.5 The Derived Interface Conjecture in Metric Terms 17

5 Synthesis: The Three-Level Chain **17**

6 Projection Chains and Composed Loss **19**
6.1 Composing Projections 19
6.2 The Distinction Functional Along a Chain 20
6.3 When Intermediate Projections Help 20

7 Conditional Distinctions and Contextual Comparison **21**
7.1 Comparing Interfaces in Context 21
7.2 The Distinction Profile 22
7.3 Mutual Distinction and Shared Information 22

8 Application Corollaries **23**
8.1 Physics 23
8.2 Cognition 23
8.3 Institutions 24

Conclusion **24**

Orientation

Every observation is a projection. Whatever generates an observation — a physical system, a cognitive process, a social institution, a measuring device — projects its internal configuration onto a reduced description that can be received, recorded, and compared. The projection is in general many-to-one. Many internal configurations produce the same observation. What does not vary across a fiber is, by definition, invisible.

This paper takes that structure seriously as a mathematical object. It does not ask what the hidden internal space is, or how it might be reconstructed. It asks what geometry is imposed on the space of observations by the existence of fibers, and what the non-invertibility of projection implies for irreversibility, memory, and the comparison of theories.

The answer, developed in three sections, is that the geometry of observation is an information geometry: the natural metric on observable interfaces is the Jensen–Shannon distance, the curvature of the resulting space measures falsifiability and opacity, and the minimum information destroyed by any observation is the reconstruction defect of the generating projection. These are not analogies between different domains. They are the same mathematical structure appearing in different applications.

The present document is the mathematical core of a larger monograph, *Distinction Geometry*. It is intended to be read first: the applications — to quantum gravity, thermodynamics, cognition, and institutional theory — are corollaries of what is proved here.

1. The Projection Setting

1.1. Histories, Projections, and Fibers

Throughout this paper, M denotes a *space of histories*: a set whose elements are complete trajectories of a system over a time interval, not instantaneous configurations. This choice is deliberate. The phenomena we wish to unify — irreversibility, memory, learning, paradigm change — are inherently historical. They concern how the past constrains the future and what of the past survives into observation. A state-based formalism would need to carry additional temporal structure explicitly; the history formalism carries it automatically.

Definition 1.1 (Projection Setting)**.** *A* projection setting *is a triple* (M, O, Π) *where* M *is a measurable space of histories,* O *is a measurable space of observable*

outcomes, and $\Pi : M \to O$ *is a surjective measurable map called the* projection. *The* fiber *over* $o \in O$ *is*

$$F_o = \Pi^{-1}(o) = \{h \in M : \Pi(h) = o\}. \tag{1}$$

Two histories $h, h' \in M$ *are* observationally equivalent*, written* $h \sim h'$*, if* $\Pi(h) = \Pi(h')$.

Remark 1.1 (On the category of M). The appropriate category for M varies by application. For quantum field theory, M is a space of distributional field configurations with the weak-* topology. For the Krein-space formulation of higher-derivative gravity, M is a Krein space (an indefinite inner-product space with a $\mathcal{K}_+ \oplus \mathcal{K}_-$ decomposition). For classical mechanics, M is a cotangent bundle. For cognitive models, M is a space of neural activation histories. For institutional models, M is a space of decision trajectories. The framework requires only that M be measurable and that Π be surjective and measurable. Additional structure — topology, metric, linear structure — may be present and will be used when available, but is not presupposed.

1.2. Admissibility and the Projection Sufficiency Principle

In applications, not every history in M is physically, cognitively, or institutionally realizable. An admissibility predicate selects the realizable ones.

Definition 1.2 (Admissibility)**.** *Given a projection setting* (M, O, Π)*, an* admissibility predicate *is a measurable function* $A_O : O \to \{0, 1\}$*. The induced predicate on* M *is* $A_M(h) = A_O(\Pi(h))$*. A history* h *is* admissible *if* $A_M(h) = 1$.

The definition places admissibility on O, not on M. This is the central methodological commitment of the framework.

Theorem 1.3 (Projection Sufficiency)**.** *If physical, cognitive, or institutional predictions depend only on* $A_O(\Pi(h))$*, then no condition on* $h \in M$ *is necessary unless it changes* $\Pi(h)$.

Proof. Let $h, h' \in M$ with $\Pi(h) = \Pi(h')$. Then

$$A_M(h) = A_O(\Pi(h)) = A_O(\Pi(h')) = A_M(h').$$

Any property that distinguishes h from h' while leaving $\Pi(h)$ unchanged has no effect on admissibility. Since all predictions are functions of admissibility and observations, and since both are determined by $\Pi(h)$, no condition on fiber structure can affect any prediction. □

Corollary 1.4 (The Injectivity Special Case)**.** *When* Π *is injective,* $\Pi(h) = \Pi(h')$ *implies* $h = h'$*, so every fiber is a singleton. There are no hidden directions in the fiber, and every internal property of* h *is directly observable. This is the condition corresponding to the Hilbert Assumption in quantum mechanics (the identification of the state space with the observable space), naïve realism in epistemology (the identification of mental content with external reality), and the legibility assumption in institutional theory (the identification of an organization's behavior with its internal dynamics).*

Remark 1.2 (Pathologies in M versus pathologies in O). A *coordinate pathology* is a property of some $h \in M$ that has no bearing on $\Pi(h)$. An *observable pathology* is a violation of admissibility in O: $A_O(\Pi(h)) = 0$. By Theorem 1.3, coordinate pathologies have no observable consequences. Whether a specific pathology in M induces an observable pathology depends on the structure of Π, not on the pathology alone. This elementary observation underlies the resolution of the ghost problem in higher-derivative quantum gravity, the resolution of the Ostrogradsky instability in gravitational cosmology, and the analysis of institutional dysfunction in large organizations that continue to function despite severe internal contradictions.

2. Reconstruction Defect and Irreversibility

2.1. The Reconstruction Problem

Given an observation $o = \Pi(h)$, can the generating history h be recovered? This is the reconstruction problem. When Π is injective, the answer is yes: the unique preimage $\Pi^{-1}(o)$ recovers h exactly. When Π is many-to-one, the fiber F_o contains multiple histories, and reconstruction must choose among them.

Definition 2.1 (Reconstruction Map and Defect)**.** *A* reconstruction map *is a measurable function* $R : O \to M$ *such that* $\Pi(R(o)) = o$ *for all* $o \in O$ *(that is,* R *is a section of* Π*). Given a metric* d *on* M*, the* reconstruction defect *of* R *at* h *is*

$$\Delta_R(h) = d\big(h,\, R(\Pi(h))\big). \tag{2}$$

The worst-case reconstruction defect *of* R *is* $\|\Delta_R\|_\infty = \sup_{h \in M} \Delta_R(h)$*. The* minimum reconstruction defect *of the projection is*

$$\delta(\Pi) = \inf_R \|\Delta_R\|_\infty, \tag{3}$$

where the infimum is taken over all reconstruction maps.

Theorem 2.2 (Irreversibility from Non-Injectivity)**.** *Let $\Pi : M \to O$ and let d be a metric on M. Then*

$$\delta(\Pi) \;\geq\; \frac{1}{2} \sup_{o \in O} \operatorname{diam}(F_o), \tag{4}$$

where $\operatorname{diam}(F_o) = \sup_{h,h' \in F_o} d(h, h')$*. In particular, if some fiber F_o has positive diameter then every reconstruction map R has positive worst-case defect on that fiber:* $\sup_{h \in F_o} \Delta_R(h) \geq \frac{1}{2} \operatorname{diam}(F_o) > 0$.

Proof. Fix any $o \in O$ and any two points $h, h' \in F_o$. For any reconstruction map R, the triangle inequality gives

$$d(h, h') \leq d(h, R(o)) + d(R(o), h') = \Delta_R(h) + \Delta_R(h').$$

Therefore $\max(\Delta_R(h), \Delta_R(h')) \geq \frac{1}{2} d(h, h')$. Taking the supremum over $h, h' \in F_o$ and then over all o and all reconstruction maps R:

$$\delta(\Pi) = \inf_R \sup_{h \in M} \Delta_R(h) \geq \frac{1}{2} \sup_{o \in O} \sup_{h,h' \in F_o} d(h, h') = \frac{1}{2} \sup_{o \in O} \operatorname{diam}(F_o).$$

When d separates points, any non-singleton fiber has positive diameter, so $\delta(\Pi) > 0$ whenever Π is non-injective. □

Remark 2.1. Theorem 2.2 is the precise mathematical content of the claim that information is lost under projection. The reconstruction defect $\delta(\Pi)$ is a measure of how much is lost: it is the minimum error that any reconstruction must make, regardless of how cleverly it tries to infer the generating history from the observation. Irreversibility is not a property of the dynamics of the system but of the projection by which the system is observed. A dynamical system whose intrinsic evolution is time-reversible appears irreversible to any observer whose projection Π is non-injective.

2.2. The Defect Spectrum

The minimum worst-case defect $\delta(\Pi)$ is a single number. It is useful to have a finer characterisation of how reconstruction defect is distributed across M.

Definition 2.3 (Defect Spectrum)**.** *Given (M, O, Π, d), the* defect spectrum *is the function $\delta_* : [0, \infty) \to [0, 1]$ defined by*

$$\delta_*(t) = \sup_R \mu\big(\{h \in M : \Delta_R(h) \leq t\}\big), \tag{5}$$

where μ is a reference probability measure on M and the supremum is over all

reconstruction maps $R : O \to M$. Thus $\delta_*(t)$ is the largest fraction of histories that the best reconstruction map can recover to within distance t.

The defect spectrum interpolates between $\delta_*(0) = 0$ (when Π is non-injective, no reconstruction map achieves zero defect everywhere) and $\delta_*(\infty) = 1$ (the entire history space is recovered to within infinite tolerance). Its shape characterises the nature of the information loss:

Proposition 2.4 (Defect Spectrum at Zero)**.** *$\delta_*(0) = 1$ if and only if Π is injective. When Π is non-injective, $\delta_*(0) < 1$: even the best reconstruction map fails on a positive-measure set of histories.*

Proof. If Π is injective, $R = \Pi^{-1}$ satisfies $\Delta_R(h) = 0$ for all h, so $\delta_*(0) = \mu(M) = 1$. If Π is non-injective, some fiber F_o is non-singleton; by Theorem 2.2, any R incurs positive defect on at least one element of F_o, so $\mu(\{h : \Delta_R(h) = 0\}) < \mu(M) = 1$ for every R, giving $\delta_*(0) < 1$. □

2.3. Residual Uncertainty and Reconstruction Defect

The information lost under projection is quantified by the *conditional entropy of h given $\Pi(h)$*:

$$H(h \mid \Pi(h)) = \sum_{o \in O} \nu(o)\, H(\mu_{F_o}), \tag{6}$$

where μ_{F_o} is the conditional distribution of h on the fiber F_o and $\nu = \Pi_*\mu$. This is the residual uncertainty about the generating history that survives after observing $\Pi(h) = o$: the average entropy within fibers, not the entropy of the projected distribution.

Remark 2.2 (Why $H(\Pi_*\mu)$ is the wrong bound). The Shannon entropy $H(\nu) = H(\Pi_*\mu)$ measures uncertainty about which fiber an observation falls in; it does not measure uncertainty about the history *within* a fiber. A projection that collapses everything into a single fiber has $H(\nu) = 0$ but $H(h \mid \Pi(h))$ arbitrarily large. The operative quantity for reconstruction error is the conditional entropy $H(h \mid \Pi(h))$, not the marginal $H(\Pi_*\mu)$.

Theorem 2.5 (Fano-Type Bound on Reconstruction)**.** *Let $(M, \mathcal{F}, \mu)$ be a probability space with M finite, and let $R : O \to M$ be any reconstruction map. Then*

$$P_e(R) \;\geq\; \frac{H(h \mid \Pi(h)) - \log 2}{\log |M|}, \tag{7}$$

where $P_e(R) = \mu(\{h : R(\Pi(h)) \neq h\})$ is the error probability. Consequently, whenever $H(h \mid \Pi(h)) > \log 2$, no reconstruction map achieves zero error probability.

Proof. Apply Fano's inequality [3] to the channel $h \to \Pi(h) \to R(\Pi(h))$. With input h, output $R(\Pi(h))$, and error event $\{R(\Pi(h)) \neq h\}$, Fano's inequality gives $H(h \mid \Pi(h)) \leq H(P_e) + P_e \log(|M| - 1) \leq \log 2 + P_e \log |M|$, which rearranges to (7). □

Remark 2.3 (Metric Defect via Rate-Distortion). For a metric lower bound on $\mathbb{E}[\Delta_R]$ rather than on error probability, rate-distortion theory gives $\mathbb{E}_\mu[\Delta_R] \geq D^*(H(h \mid \Pi(h)))$, where $D^*(I)$ is the rate-distortion function of the fiber-conditional distribution at information constraint I [3]. In both cases the operative quantity is $H(h \mid \Pi(h))$ (6).

Corollary 2.6 (Thermodynamic Interpretation)**.** *When M is the microstate space of a thermodynamic system, O is the macrostate space, and Π is the standard coarse-graining map, the Boltzmann entropy $S = k_B \log |F_o|$ (for uniform measure on M) is a measure of reconstruction defect: the logarithm of the number of microstates compatible with the macrostate is the log-volume of the fiber. The second law of thermodynamics, under this reading, is the statement that a system initialized in a small fiber and evolving under a measure-preserving dynamics will, under coarse-graining, appear to move toward states with larger fibers, increasing the reconstruction defect.*

2.4. Irreversibility as Accumulated Defect

For systems with a dynamical evolution, irreversibility is not merely the existence of reconstruction defect but its persistence and growth under the dynamics.

Definition 2.7 (Dynamical Irreversibility)**.** *Let $\phi_t : M \to M$ be a flow on M (the dynamics), and let Π be the observation projection. The* accumulated reconstruction defect *at time T starting from h is*

$$\mathcal{I}(h, T) = \inf_R \int_0^T \Delta_R(\phi_t(h))\, dt. \tag{8}$$

A system is dynamically irreversible at h *if $\mathcal{I}(h, T) > 0$ for all $T > 0$ and all reconstruction maps R.*

Proposition 2.8 (Conditions for Dynamical Irreversibility)**.** *If $\Pi \circ \phi_t$ is non-injective for μ-almost every $t > 0$, then the system is dynamically irreversible at μ-almost every $h \in M$.*

Proof. Non-injectivity of $\Pi \circ \phi_t$ means there exist $h \neq h'$ with $\Pi(\phi_t(h)) = \Pi(\phi_t(h'))$. By Theorem 2.2, any reconstruction map incurs positive defect at either $\phi_t(h)$ or $\phi_t(h')$ for each such t. Integrating over t gives positive accumulated defect. □

Remark 2.4 (The Arrow of Time). The arrow of time, in the framework of reconstruction defect, is the direction in which $\mathcal{I}(h, T)$ grows. It is not a property of the fundamental dynamics ϕ_t — which may be time-reversible at the level of M — but of the projection Π under which the system is observed. A perfectly reversible dynamics observed through a non-injective projection appears irreversible because the accumulated defect grows with T: the observer loses the ability to determine which of the histories compatible with the current observation was the actual generating history. This is the precise sense in which the arrow of time is a projection artifact.

3. The Distinction Functional

3.1. The Problem of Comparing Interfaces

Two observers equipped with different projections Π_1 and Π_2 will assign different probability distributions to the same underlying space of histories. How should the difference between these distributions be measured?

The answer must satisfy several constraints. First, it must be *interface-native*: the comparison should require no reference to the underlying space M or the projections Π_1, Π_2. Once the distributions $P = \Pi_{1*}\mu_1$ and $Q = \Pi_{2*}\mu_2$ on O are given, the comparison should proceed entirely within O.

Second, the measure should be *symmetric*: the distance from P to Q should equal the distance from Q to P. There is no privileged interface.

Third, the measure should be *operationally meaningful*: it should have an interpretation in terms of distinguishability, information, or some quantity that can in principle be estimated from observations.

Fourth, it should be *a metric*: zero if and only if $P = Q$, satisfying the triangle inequality.

The Jensen–Shannon divergence satisfies all four constraints. We develop its properties systematically in this section.

3.2. Definition and Basic Properties

Definition 3.1 (Jensen–Shannon Divergence and Distance)**.** *Let P and Q be probability distributions on a measurable space O. Let $M_{PQ} = \frac{1}{2}(P + Q)$ be their equal-weight mixture. The* Jensen–Shannon divergence *is*

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M_{PQ}) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M_{PQ}), \tag{9}$$

where $D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\log(P(x)/Q(x))$ *is the Kullback–Leibler divergence. Equivalently,*

$$\mathrm{JSD}(P\|Q) = H(M_{PQ}) - \frac{1}{2}\big(H(P) + H(Q)\big), \tag{10}$$

where $H(\cdot)$ *is the Shannon entropy. The* Jensen–Shannon distance *is* $d_{\mathrm{int}}(P,Q) = \mathrm{JSD}^{1/2}(P\|Q)$.

Proposition 3.2 (Metric Properties)**.** *The Jensen–Shannon distance* d_{int} *is a metric on the space of probability distributions on* O*. Specifically:*

(i) $d_{\mathrm{int}}(P,Q) \geq 0$*, with equality if and only if* $P = Q$*;*

(ii) $d_{\mathrm{int}}(P,Q) = d_{\mathrm{int}}(Q,P)$*;*

(iii) $d_{\mathrm{int}}(P,R) \leq d_{\mathrm{int}}(P,Q) + d_{\mathrm{int}}(Q,R)$.

Proof. Properties (i) and (ii) follow immediately from the definition. $\mathrm{JSD}(P\|Q) = 0$ iff both KL terms vanish, iff $P = M_{PQ} = Q$. Symmetry is manifest from the equal-weight mixture. The triangle inequality for $d_{\mathrm{int}} = \mathrm{JSD}^{1/2}$ was established by Endres and Schindelin [1] using the concavity of $\sqrt{\cdot}$ and the properties of KL divergence. The key step is that $\mathrm{JSD}^{1/2}$ satisfies the triangle inequality while JSD alone does not. □

Remark 3.1. The need to take the square root — that $\mathrm{JSD}^{1/2}$ is a metric while JSD is not — is the information-geometric signature of the curvature we will compute in Section 4. In a flat information geometry, the square root would be unnecessary; the need for it indicates that the space of probability distributions is positively curved in the Fisher–Rao metric.

3.3. The Distinction Functional

Definition 3.3 (Distinction Functional)**.** *The* distinction functional *between observable interfaces* P *and* Q *is*

$$\mathcal{D}(P,Q) = H(M_{PQ}) - \frac{1}{2}\big(H(P) + H(Q)\big) = \mathrm{JSD}(P\|Q). \tag{11}$$

The entropy decomposition (10) gives each term a precise operational meaning, and their combination characterises the information carried by the distinction between P and Q.

Theorem 3.4 (Operational Meaning of the Distinction Functional)**.** *The distinction functional* $\mathcal{D}(P,Q)$ *measures the expected reduction in uncertainty about which interface generated an observation, given that the observation has been made.*

Equivalently, it is the information that is destroyed when the distinction between P and Q is forgotten by replacing both with their mixture M_{PQ}.

Proof. Consider an observer who knows an observation $x \in O$ was generated by one of two sources, each equally likely: source 1 using distribution P, source 2 using distribution Q. The prior uncertainty about the source is $H(\text{source}) = \log 2$. The observation x is drawn from the mixture M_{PQ}. After seeing x, the posterior uncertainty about the source is $H(\text{source} \mid x) = -\sum_{s\in\{1,2\}} P(s \mid x) \log P(s \mid x)$. The expected posterior uncertainty over observations is $\sum_x M_{PQ}(x) H(\text{source} \mid x)$.

By the chain rule of entropy applied to the joint (x, source):

$$H(x, \text{source}) = H(\text{source}) + H(x \mid \text{source}) = \log 2 + \frac{1}{2}(H(P) + H(Q)),$$
$$H(x, \text{source}) = H(x) + H(\text{source} \mid x) = H(M_{PQ}) + H(\text{source} \mid x).$$

Therefore $H(\text{source} \mid x) = \log 2 + \frac{1}{2}(H(P) + H(Q)) - H(M_{PQ}) = \log 2 - \mathcal{D}(P, Q)$. The information gain from observing x is $\log 2 - H(\text{source} \mid x) = \mathcal{D}(P, Q)$. When the distinction is forgotten (P and Q are merged into M_{PQ}), this information gain vanishes: the expected uncertainty about the source becomes $\log 2$ (no information about the source), and the distinction functional equals exactly the information that was lost. □

Corollary 3.5 (Monotonicity and Data Processing)**.** *The distinction functional satisfies the data processing inequality: for any measurable map $f : O \to O'$,*

$$\mathcal{D}(f_* P,\ f_* Q) \leq \mathcal{D}(P, Q). \tag{12}$$

Any further processing of observations can only reduce the distinction between the interfaces.

Proof. This follows from the data processing inequality for KL divergence:

$$D_{\mathrm{KL}}(f_* P \| f_* M_{PQ}) \leq D_{\mathrm{KL}}(P \| M_{PQ}),$$

and similarly for Q. □

3.4. Interface-Nativity

Proposition 3.6 (Interface-Nativity)**.** *The distinction functional $\mathcal{D}(P, Q)$ is interface-native: its definition and computation require no reference to any underlying space M or projection Π.*

Proof. Definition 3.3 uses only the distributions P and Q on the observable space O, their mixture M_{PQ}, and the Shannon entropy H. No element of M appears, no projection Π is referenced, and no admissibility manifold is invoked. The computation can be performed by any agent who has access to O and can sample from P and Q, without knowledge of how those distributions were generated. □

Remark 3.2 (Contrast with the Inf-Sup Formulation). An alternative distance between theories with underlying spaces $M^{(1)}$ and $M^{(2)}$ would be

$$d_{\text{ext}}(M^{(1)}, M^{(2)}) = \inf_{\phi} \sup_{h} \big\| \Pi_1(h) - \Pi_2(\phi(h)) \big\|_O, \tag{13}$$

where the infimum runs over admissibility-preserving maps ϕ. This formulation is not interface-native: the infimum over ϕ requires access to $M^{(1)}$ and $M^{(2)}$. It frames the question as "how similar can the internal structures be made to look?" rather than "how similar are the observable outputs?" The distinction functional answers the second question without invoking the first, and is therefore the appropriate primitive for an interface-native theory.

3.5. Gauge Orbits and Hidden Directions

Theorem 3.7 (Gauge Orbits as Zero-Distinction Fibers)**.** *Let (M, O, Π) be a projection setting. The observational equivalence relation $h \sim h'$ (Definition 1.1) is the zero-level set of the induced distinction functional between the Dirac measures at the projected points:*

$$h \sim h' \quad \Longleftrightarrow \quad \mathcal{D}(\delta_{\Pi(h)},\, \delta_{\Pi(h')}) = 0, \tag{14}$$

where δ_o denotes the Dirac measure concentrated at $o \in O$.

Proof. The distinction functional compares probability distributions on O. For point masses, $\mathcal{D}(\delta_o, \delta_{o'}) = \mathrm{JSD}(\delta_o \| \delta_{o'})$. Since $H(\delta_o) = 0$ for any point mass and $H(\frac{1}{2}(\delta_o + \delta_{o'})) = \log 2$ when $o \neq o'$ and 0 when $o = o'$, we get $\mathcal{D}(\delta_o, \delta_{o'}) = \log 2$ when $o \neq o'$ and 0 when $o = o'$. Therefore

$$\mathcal{D}(\delta_{\Pi(h)},\, \delta_{\Pi(h')}) = 0 \iff \Pi(h) = \Pi(h'),$$

which is the definition of $h \sim h'$. Gauge orbits are the fibers $F_o = \Pi^{-1}(o)$: the zero-distinction equivalence classes. □

Definition 3.8 (Hidden Admissibility Direction)**.** *Let h_ϵ be a smooth curve in M*

with $h_0 = h$. *The* first-order observable variation *is*

$$\delta_{\text{obs}}(h, \dot{h}) = \frac{d}{d\epsilon} d_{\text{int}}\big(\Pi(h),\, \Pi(h_\epsilon)\big)\Big|_{\epsilon=0}. \tag{15}$$

The direction $\dot{h} = \frac{dh_\epsilon}{d\epsilon}|_{\epsilon=0}$ *is* hidden to first order *if* $\delta_{\text{obs}}(h, \dot{h}) = 0$. *It is* fully hidden *if* $\mathcal{D}(\delta_{\Pi(h)}, \delta_{\Pi(h_\epsilon)}) = 0$ *for all sufficiently small* $\epsilon > 0$.

Example 3.1 (Ghost States as Hidden Directions). In the Krein-space formulation of higher-derivative quantum gravity, $M = \mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-$ and Π maps each state to its observable transition probabilities via $P(F \mid I) = \text{tr}_+((\Pi_F S \Pi_I)^\dagger (\Pi_F S \Pi_I))$. A ghost state $g \in \mathcal{K}_-$ generates a deformation $h_\epsilon = h + \epsilon g$ in M. If ghost-parity symmetry holds ($[G, S] = 0$), then the contribution of g to any observable transition probability cancels in the trace over $\mathcal{K}$, so $\Pi(h + \epsilon g) = \Pi(h)$ for all ϵ. The deformation in the g-direction is fully hidden: the ghost is a hidden admissibility direction with $\mathcal{D}(\Pi(h), \Pi(h + \epsilon g)) = 0$ for all ϵ. Whether ghost-parity symmetry holds in the full spin-2 sector of quadratic gravity remains open; the present framework identifies this as precisely the question of whether the ghost direction is fully hidden.

Example 3.2 (Gauge Transformations as Hidden Directions). In Yang–Mills theory, M is the space of gauge field configurations A_μ and Π is the map to gauge-invariant observables (Wilson loops, field strength curvature, etc.). A gauge transformation $A_\mu \mapsto A_\mu + D_\mu \lambda$ is a deformation along a hidden direction: by gauge invariance, $\Pi(A_\mu + D_\mu \lambda) = \Pi(A_\mu)$, so the deformation is fully hidden. The gauge orbit through A_μ is the fiber $F_{\Pi(A_\mu)}$.

Example 3.3 (Cognitive Hidden Directions). For a cognitive agent with internal representation space M and perceptual projection Π to sensory observations, a hidden direction is an internal change that leaves all perceptual outputs unchanged. This includes: changes in sub-threshold neural activity, shifts in attentional allocation that do not affect behavioral output, alterations in implicit beliefs that have not yet influenced action, and unconscious emotional states that are not expressed in observable behavior. The distinction functional $\mathcal{D}(\delta_{\Pi(h)}, \delta_{\Pi(h+\epsilon\delta h)})$ measures whether the internal change δh is cognitively legible to an external observer.

4. The Geometry of Interface Space

4.1. The Space of Theories as a Metric Space

Let $\mathcal{M}$ denote the space of all observable interfaces: all probability distributions that can be generated by some admissible projection setting (M, O, Π, μ). Two inter-

faces $P, Q \in \mathcal{M}$ are compared by the distinction functional $\mathcal{D}(P, Q) = \mathrm{JSD}(P\|Q)$, and the interface distance $d_{\mathrm{int}} = \mathcal{D}^{1/2}$ makes $\mathcal{M}$ a metric space by Proposition 3.2.

Definition 4.1 (Theory Space)**.** *The* theory space *of a given observable domain O is the metric space $(\mathcal{M}, d_{\mathrm{int}})$ of probability distributions on O, equipped with the Jensen–Shannon distance.*

Remark 4.1. The elements of $\mathcal{M}$ are observable distributions, not internal models. Two theories with very different internal structures — different M's, different Π's, different dynamics on M — may occupy the same point in $(\mathcal{M}, d_{\mathrm{int}})$ if their projected distributions agree. The theory space is not a space of theories in the traditional sense (axiomatic systems, Lagrangians, Hamiltonians) but a space of what theories produce: observable distributions over outcomes. It is the appropriate space for empirical comparison.

4.2. Local Riemannian Structure: The Fisher Metric

For a parametric family of distributions $\{P_\lambda\}_{\lambda \in \Lambda}$, the distinction functional induces a Riemannian metric on Λ.

Theorem 4.2 (Fisher Metric from JSD)**.** *Let $\lambda \mapsto P_\lambda$ be a smooth map from an open subset $\Lambda \subseteq \mathbb{R}^n$ to probability distributions on O. Then the Jensen–Shannon divergence satisfies*

$$\mathrm{JSD}(P_\lambda,\, P_{\lambda+\epsilon}) = \frac{1}{8}\, g^F_{ij}(\lambda)\, \epsilon^i \epsilon^j + O(|\epsilon|^3), \tag{16}$$

so that $d^2_{\mathrm{int}} = \mathrm{JSD} = \frac{1}{8} g^F_{ij} \epsilon^i \epsilon^j + O(|\epsilon|^3)$, where

$$g^F_{ij}(\lambda) = \sum_{x \in O} P_\lambda(x)\, \frac{\partial \log P_\lambda(x)}{\partial \lambda^i}\, \frac{\partial \log P_\lambda(x)}{\partial \lambda^j} \tag{17}$$

is the Fisher information metric on Λ.

Proof. Let $\delta P(x) = P_{\lambda+\epsilon}(x) - P_\lambda(x) = \partial_i P_\lambda(x)\, \epsilon^i + O(|\epsilon|^2)$ and $M(x) = P_\lambda(x) + \frac{1}{2}\delta P(x)$. Expanding $D_{\mathrm{KL}}(P_\lambda \| M)$:

$$D_{\mathrm{KL}}(P_\lambda \| M) = \sum_x P_\lambda(x) \log \frac{P_\lambda(x)}{M(x)} = \sum_x P_\lambda(x) \log\left(1 - \frac{\delta P(x)}{2M(x)}\right).$$

Since $M(x) = P_\lambda(x)(1 + O(|\epsilon|))$, we have $\frac{\delta P(x)}{2M(x)} = \frac{\delta P(x)}{2P_\lambda(x)} + O(|\epsilon|^2)$. Using $\log(1 - u) = -u - u^2/2 + O(u^3)$:

$$D_{\mathrm{KL}}(P_\lambda \| M) = \frac{1}{8} \sum_x \frac{(\delta P(x))^2}{P_\lambda(x)} + O(|\epsilon|^3).$$

By symmetry $D_{\mathrm{KL}}(P_{\lambda+\epsilon}\|M)$ contributes the same leading term. Combining both KL terms (JSD $= \frac{1}{2}D_{\mathrm{KL}}(P_\lambda\|M) + \frac{1}{2}D_{\mathrm{KL}}(P_{\lambda+\epsilon}\|M)$), each contributing $\frac{1}{8}\sum_x(\delta P)^2/P_\lambda$ at leading order:

$$\mathrm{JSD}(P_\lambda,\, P_{\lambda+\epsilon}) = \frac{1}{8}\sum_x \frac{(\delta P(x))^2}{P_\lambda(x)} + O(|\epsilon|^3) = \frac{1}{8}\, g^F_{ij}(\lambda)\, \epsilon^i \epsilon^j + O(|\epsilon|^3),$$

where $g^F_{ij} = \sum_x P_\lambda(x)(\partial_i \log P_\lambda)(\partial_j \log P_\lambda)$ is the Fisher information matrix. Since $d_{\mathrm{int}} = \mathrm{JSD}^{1/2}$, we have $d^2_{\mathrm{int}} = \mathrm{JSD} = \frac{1}{8}g^F_{ij}\epsilon^i\epsilon^j + O(|\epsilon|^3)$, confirming (16). □

Remark 4.2 (Relationship to Information Geometry). Theorem 4.2 establishes that $(\mathcal{M}, d_{\mathrm{int}})$ is locally a statistical manifold [2] with the Fisher–Rao metric. The factor $\frac{1}{8}$ is a convention-dependent rescaling (it becomes $\frac{1}{2}$ in the natural units of information geometry where logarithms are taken base e). The key point is that the information geometry of the space of theories is determined entirely by the Fisher information of the observable distributions, without reference to the internal models that generate them.

4.3. Curvature and Its Interpretation

The Riemannian curvature of $(\mathcal{M}, g^F)$ is the standard Riemannian curvature of the statistical manifold with Fisher metric. It has been computed explicitly for many exponential families [2]. Here we give its operational interpretation in the theory-comparison setting.

Definition 4.3 (Sectional Curvature of Theory Space)**.** *The* sectional curvature $\kappa(\sigma)$ *of* $(\mathcal{M}, g^F)$ *at a point* P_λ *in a 2-plane* $\sigma \subset T_{P_\lambda}\mathcal{M}$ *is the Gaussian curvature of the 2-dimensional surface swept out by geodesics from* P_λ *in directions* σ*.*

Theorem 4.4 (Operational Interpretation of Curvature)**.** *At a theory* $P_\lambda \in \mathcal{M}$*:*

(i) Positive sectional curvature in direction σ *means that theories near* P_λ *in the* σ*-direction are more similar in their observable predictions than their parameter-space distance suggests. Geodesic spheres of radius* r *in theory space have* smaller *volume than Euclidean spheres of radius* r *(the Jacobi fields converge), so more parameter-space directions map into the same observable neighbourhood. This is the geometry of* overfitting*: many internal models produce observationally indistinguishable predictions.*

(ii) Negative sectional curvature in direction σ *means that theories near* P_λ *in the* σ*-direction diverge in their observable predictions faster than their parameter-space distance suggests. Geodesic spheres of radius* r *have* larger

volume than Euclidean spheres of radius r (the Jacobi fields diverge), so small parameter changes produce large observable changes. This is the geometry of high falsifiability*: small changes in the theory produce large changes in predictions.*

(iii) Zero sectional curvature is the Euclidean regime: parameter-space distance and observable-prediction distance scale proportionally, and the theory is locally well-identified.

Proof of (i) and (ii). The Jacobi equation governs the evolution of nearby geodesics: if $J(t)$ is a Jacobi field along a unit-speed geodesic $\gamma(t)$, then $\|J(t)\| \approx t - \frac{\kappa}{6}t^3 + O(t^5)$ for small t, where κ is the sectional curvature in the plane $\{\dot{\gamma}, J(0)\}$. When $\kappa > 0$, $\|J(t)\| < t$: nearby geodesics are closer than linearly predicted, meaning theories at parameter distance t are d_{int}-closer than expected, i.e., more observationally similar. When $\kappa < 0$, $\|J(t)\| > t$: theories at parameter distance t are more observationally different than the Euclidean approximation predicts, i.e., more falsifiable. □

Example 4.1 (Curvature in Specific Families). For exponential family distributions $P_\lambda(x) = \exp(\lambda \cdot T(x) - A(\lambda))$ with sufficient statistic T, the Fisher metric is $g^F_{ij} = \partial_i \partial_j A(\lambda)$ (the Hessian of the log-partition function). The curvature is determined by third-order cumulants of T under P_λ. In the Gaussian family $\mathcal{N}(\mu, \sigma^2)$, the sectional curvature in the $(\mu, \log \sigma)$ plane is $-\frac{1}{2}$: the space of Gaussian distributions is a hyperbolic space. This means distinguishing Gaussian theories by their observable predictions is easier than parameter-space distance would suggest — small changes in (μ, σ) produce large changes in d_{int}.

4.4. Geodesics as Paths of Minimal Accumulated Distinction

Definition 4.5 (Geodesic in Theory Space)**.** *A* geodesic *in $(\mathcal{M}, g^F)$ is a smooth curve $\lambda(t)$ in the theory parameter space that satisfies the geodesic equation $\nabla_{\dot{\lambda}}\dot{\lambda} = 0$ with respect to the Levi-Civita connection of the Fisher metric. Its interpretation: the geodesic between theories P and Q is the path through theory space that minimises the total accumulated distinction between successive theories along the path,*

$$\int_0^1 d_{\text{int}}\big(P_{\lambda(t)}, P_{\lambda(t+dt)}\big) = \int_0^1 \sqrt{g^F_{ij}\dot{\lambda}^i\dot{\lambda}^j}\, dt. \tag{18}$$

Remark 4.3 (Scientific and Cognitive Interpretation). The geodesic distance $d_{\text{int}}(P, Q)$ between two theories is the minimum total accumulated distinction along any path connecting them. In the context of scientific theory change, a

paradigm shift from theory P to theory Q that follows the geodesic is the shift that introduces the minimum total observational difference at each step. A shift that follows a longer path introduces unnecessary intermediate distinctions — it makes claims that cannot be verified by any available observations. The geodesic is the *most parsimonious* path between two empirical positions. Occam's razor, in this framework, is the injunction to travel geodesically.

4.5. The Derived Interface Conjecture in Metric Terms

The physical application of theory space geometry provides a precise formulation of what it means for two theories of gravity to make the same observable predictions below the Planck scale.

Definition 4.6 (Sub-Planckian Interface Equivalence)**.** *Two quantum gravity theories T_1 and T_2 are* sub-Planckian interface equivalent *if $d_{\text{int}}(P_1^{(E)}, P_2^{(E)}) = 0$ for all observable energies $E \ll M_{\text{Pl}}$, where $P_i^{(E)}$ is the distribution of scattering outcomes at energy E predicted by T_i.*

Principle 4.1 (Derived Interface Conjecture, metric form)**.** Hilbert-space quantum gravity and Krein-space quadratic gravity are sub-Planckian interface equivalent: they satisfy $d_{\text{int}} = 0$ at all energies accessible to current and foreseeable experiment. Above the Planck scale, they diverge: $d_{\text{int}} > 0$, with the distance controlled by the mass of the spin-2 ghost $m_2 = M_{\text{Pl}}/\sqrt{2\beta}$.

The conjecture is falsifiable in the precise sense of Theorem 4.4: it predicts that the theory space curvature in the direction separating the two theories is controlled by the energy scale at which they diverge, and that no sub-Planckian measurement can distinguish them.

5. Synthesis: The Three-Level Chain

The three structures developed in Sections 2–4 form a coherent chain:

$$M \xrightarrow{\Pi} O \xrightarrow{\mathcal{D}} (\mathcal{M},\, d_{\text{int}}). \tag{19}$$

The first arrow discards internal structure: what survives is the observable interface $P = \Pi_* \mu \in \mathcal{M}$. The information lost in this step is measured by the reconstruction defect $\delta(\Pi)$ of Section 2. The second arrow measures distinguishability between interfaces: what survives is a geometry on the space of theories. The information destroyed when the distinction between two interfaces is forgotten is measured by the distinction functional $\mathcal{D}$ of Section 3.

The central question shifts accordingly at each level:

Level 1 (internal):	What structures exist in M?
Level 2 (interface):	What can be inferred about M from O?
Level 3 (geometric):	What geometry do the limitations of inference impose on $\mathcal{M}$?

This paper concerns primarily Levels 2 and 3. Level 1 is the domain of specific physical, cognitive, or institutional theories; the present framework takes whatever is at Level 1 and asks what it produces and what structure that production imposes on the space of theories.

Theorem 5.1 (Unified Translation)**.** *Under the chain (19), the following translations hold:*

(i) A gauge orbit is a zero-$\mathcal{D}$ fiber of Π (Theorem 3.7).

(ii) A hidden admissibility direction is a tangent direction to M along which $d_{\text{int}}(\Pi(h), \Pi(h_\epsilon)) = 0$ to first order (Definition 3.8).

(iii) Irreversibility is persistent positive reconstruction defect $\Delta_R > 0$ under any reconstruction map (Theorem 2.2).

(iv) The Born rule is the projection $re^{i\theta} \mapsto r^2$: the map from complex amplitude space to observable intensity space that discards the phase (a hidden direction in the fiber of the squaring map).

(v) Theory comparison is distance in $(\mathcal{M}, d_{\text{int}})$ (Definition 4.1).

(vi) The curvature of theory space is the curvature of distinguishability (Theorem 4.4).

Remark 5.1 (On the Ontological Question). The chain (19) does not answer the question "what exists?" It dissolves it. The question "what exists in M?" is replaced by "what survives projection?" (Level 2) and "what structure does the pattern of survival impose on $\mathcal{M}$?" (Level 3). Whatever exists in M — fields, states, decision trajectories, neural activations, institutional dynamics — is relevant to inquiry only to the degree that it produces distinguishable differences in O. The geometry of $(\mathcal{M}, d_{\text{int}})$ is the geometry of what matters empirically, which may be a strict subset of what exists internally.

This is not instrumentalism. It is not the claim that internal structure is unreal. It is the claim that the epistemologically primary object — the object about which

rational inquiry can make progress — is the geometry of distinguishability, not the geometry of M. Internal structure is a tool for generating and explaining observable distinctions, not an object of inquiry in its own right.

6. Projection Chains and Composed Loss

6.1. Composing Projections

Many systems involve not a single projection but a chain of projections: an internal space is projected onto an intermediate space, which is then projected again onto the final observable space. Brains project sensory signals onto perceptual representations, which are then projected onto behavioral outputs. Scientific instruments project physical quantities onto measurement readings, which are then projected onto reported values. Institutions project individual decisions onto collective actions, which are then projected onto public outcomes.

Definition 6.1 (Projection Chain)**.** *A* projection chain *of length n is a sequence of projection settings*

$$M_0 \xrightarrow{\Pi_1} M_1 \xrightarrow{\Pi_2} M_2 \;\cdots\; \xrightarrow{\Pi_n} M_n = O, \tag{20}$$

with composite projection $\Pi = \Pi_n \circ \cdots \circ \Pi_1 : M_0 \to O$. The intermediate observable spaces *are $M_1, \ldots, M_{n-1}$.*

Theorem 6.2 (Defect Accumulates Along Chains)**.** *Let $\Pi = \Pi_2 \circ \Pi_1$ be a composed projection. Then for any metrics d_0 on M_0 and d_1 on M_1:*

$$\delta(\Pi) \geq \delta(\Pi_1). \tag{21}$$

Composing with a further projection Π_2 cannot decrease the minimum reconstruction defect.

Proof. Let $R : O \to M_0$ be any reconstruction map for Π. Consider the reconstruction map $R_1 = \Pi_1 \circ R$ for Π_2: it maps $o \in O$ to $\Pi_1(R(o)) \in M_1$. For any $h \in M_0$,

$$\Delta_{R_1}(\Pi_1(h)) = d_1(\Pi_1(h),\, R_1(\Pi_2(\Pi_1(h)))) = d_1(\Pi_1(h),\, \Pi_1(R(\Pi(h)))).$$

If d_1 satisfies $d_1(\Pi_1(h), \Pi_1(h')) \leq d_0(h, h')$ (i.e., Π_1 is 1-Lipschitz), then

$$\Delta_{R_1}(\Pi_1(h)) \leq d_0(h,\, R(\Pi(h))) = \Delta_R(h).$$

Taking inf over R and sup over h:

$$\delta(\Pi_2) \leq \delta(\Pi),$$

meaning the defect of the intermediate space M_1 is no greater than that of the full chain. Equivalently, $\delta(\Pi) \geq \delta(\Pi_1)$. □

Corollary 6.3 (Monotonicity of Defect)**.** *In a projection chain $M_0 \to M_1 \to \cdots \to O$, the minimum reconstruction defect is non-decreasing: each further projection can only increase or maintain the defect. Information is monotonically lost along any chain.*

This is the reconstruction-defect version of the data processing inequality. It establishes irreversibility as monotone along projection chains: no intermediate step can recover information lost at an earlier step.

6.2. The Distinction Functional Along a Chain

Proposition 6.4 (Distinction Decreases Along Chains)**.** *For a projection chain $\Pi = \Pi_2 \circ \Pi_1$, the distinction between any two histories can only decrease:*

$$\mathcal{D}(\delta_{\Pi(h)}, \delta_{\Pi(h')}) \leq \mathcal{D}(\delta_{\Pi_1(h)}, \delta_{\Pi_1(h')}) \leq \mathcal{D}(\delta_h, \delta_{h'}), \tag{22}$$

where point masses $\delta_h, \delta_{h'}$ are the degenerate distributions on M_0.

Proof. This follows from the data processing inequality for JSD (Corollary 3.5): each further projection Π_i is a measurable map applied to the distributions, and data processing ensures $\mathcal{D}(f_*P, f_*Q) \leq \mathcal{D}(P, Q)$. □

The proposition formalises the intuition that passing information through more processing steps can only blur the distinction between its sources. A signal that is clearly distinguishable at the physical level may become indistinguishable after enough stages of perceptual, cognitive, or institutional processing.

6.3. When Intermediate Projections Help

Theorem 6.2 establishes that intermediate projections cannot decrease defect. But they can *re-structure* it: an intermediate projection Π_1 may create a representation M_1 from which subsequent reconstruction $R_1 : O \to M_1$ is more efficient than direct reconstruction $R : O \to M_0$, even though the total defect cannot decrease.

This is the formal account of *representation learning*: the task of finding an intermediate projection Π_1 such that the downstream tasks (encoded in Π_2) can be

performed with lower defect from M_1 than from M_0 directly. Good representations are intermediate projections that preserve the distinctions needed for downstream tasks while collapsing distinctions that are irrelevant to them.

Definition 6.5 (Task-Relevant Distinction Preservation)**.** *Given a downstream projection $\Pi_2 : M_1 \to O$, an intermediate projection $\Pi_1 : M_0 \to M_1$ is* task-relevant *if $\delta(\Pi_2 \circ \Pi_1) = \delta(\Pi_2)$: the composition has the same defect as the downstream projection alone. In other words, Π_1 preserves all distinctions that matter for Π_2.*

Task-relevant projections are the ideal representations: they discard exactly the information that is irrelevant to the downstream task and preserve exactly the information that is relevant. The theory of task-relevant projections is the foundation of the theory of abstraction, compression, and feature learning.

7. Conditional Distinctions and Contextual Comparison

7.1. Comparing Interfaces in Context

The distinction functional $\mathcal{D}(P, Q)$ compares two interfaces globally: it asks how different the full distributions P and Q are. But in many applications, what matters is not the global difference between two interfaces but their difference *in a particular context*: given that an observation falls in category C, how different are the two interfaces?

Definition 7.1 (Conditional Interface)**.** *Given observable distributions P and Q on O and a measurable subset $C \subseteq O$ with $P(C) > 0$ and $Q(C) > 0$, the* conditional interfaces *are*

$$P|_C(x) = \frac{P(x)\,\mathbf{1}_{x \in C}}{P(C)}, \qquad Q|_C(x) = \frac{Q(x)\,\mathbf{1}_{x \in C}}{Q(C)}. \tag{23}$$

The conditional distinction functional *is*

$$\mathcal{D}_C(P, Q) = \mathcal{D}(P|_C,\, Q|_C) = \mathrm{JSD}(P|_C \| Q|_C). \tag{24}$$

Proposition 7.2 (Properties of Conditional Distinction)**.** (i) *$\mathcal{D}_C(P, Q) = 0$ iff $P|_C = Q|_C$: the interfaces coincide within C.*

(ii) *$\mathcal{D}_C(P, Q) \leq \log 2$, with equality iff $P|_C$ and $Q|_C$ have disjoint support.*

(iii) *$\mathcal{D}_C(P, Q)$ may exceed or fall below $\mathcal{D}(P, Q)$; conditioning can reveal or suppress distinctions.*

Proof. Parts (i) and (ii) follow immediately from the corresponding properties of JSD: $\mathrm{JSD}(P\|Q) = 0 \iff P = Q$, and $\mathrm{JSD}(P\|Q) \leq \log 2$ with equality iff P and Q have disjoint support. Part (iii) follows by example: if P and Q agree on C but differ outside C, then $\mathcal{D}_C(P,Q) = 0 < \mathcal{D}(P,Q)$; if P and Q agree outside C but differ maximally within C, then $\mathcal{D}_C(P,Q) > \mathcal{D}(P,Q)$. □

7.2. The Distinction Profile

Rather than a single global comparison, the *distinction profile* of two interfaces records how their distinction varies across contexts.

Definition 7.3 (Distinction Profile)**.** *Given a partition $\{C_1, \ldots, C_k\}$ of O and distributions P, Q on O, the* distinction profile *is the vector*

$$\mathcal{D}_{[C]}(P,Q) = \big(\mathcal{D}_{C_1}(P,Q),\, \mathcal{D}_{C_2}(P,Q),\, \ldots,\, \mathcal{D}_{C_k}(P,Q)\big) \in [0, \log 2]^k. \tag{25}$$

The distinction profile records where two theories agree and where they disagree, at the resolution of the partition $\{C_i\}$. Two theories with the same global distinction $\mathcal{D}(P,Q)$ may have very different distinction profiles: one pair may disagree uniformly across all contexts, while another may agree in most contexts but disagree sharply in a few.

The distinction profile is the basis for a more refined theory of falsifiability: rather than asking "is this theory falsifiable?" (a global question), one asks "in which contexts does this prediction differ from its alternatives?" (a context-specific question). Strong falsifiability is a sharply peaked distinction profile: the two theories differ significantly in a specific, experimentally accessible context.

7.3. Mutual Distinction and Shared Information

Definition 7.4 (Mutual Distinction)**.** *Given three distributions P, Q, and R on O, the* mutual distinction *between P and Q given R is*

$$\mathcal{D}(P, Q \mid R) = \mathbb{E}_{x\sim R}\big[\mathcal{D}(\delta_{P(x)},\, \delta_{Q(x)})\big]\,, \tag{26}$$

where δ_p is the Bernoulli distribution with parameter p. This measures how much the distinction between P and Q is visible to a reference observer using distribution R.

When $R = M_{PQ} = \frac{1}{2}(P+Q)$, the mutual distinction reduces to the distinction functional $\mathcal{D}(P,Q)$. When R is the distribution of an external observer with different priors or different resolution, the mutual distinction measures the distinction

between P and Q as seen by that observer. This formulation is useful for modeling partial observability and for the theory of coordinated inquiry: two agents with different reference distributions R_1 and R_2 may perceive the distinction between two theories very differently, even when observing the same outcomes.

8. Application Corollaries

The three sections above (ğ2 on reconstruction defect, ğ3 on the distinction functional, and ğ4 on interface geometry) are the mathematical core. The following corollaries summarise their application to the three main domains of the monograph: physics, cognition, and institutions. Detailed treatments appear in the corresponding parts of *Foundations of Distinction Geometry*.

8.1. Physics

Corollary 8.1 (Ghost States)**.** *In a Krein-space quantum field theory, a ghost state $g \in \mathcal{K}_-$ is a hidden admissibility direction (Definition 3.8) if and only if $\mathcal{D}(\delta_{\Pi(h)}, \delta_{\Pi(h+\epsilon g)}) = 0$ for all sufficiently small ϵ. Whether this condition holds in the spin-2 sector of quadratic gravity is the central open problem of the programme.*

Corollary 8.2 (Thermodynamic Arrow)**.** *The thermodynamic arrow of time is the direction in which the minimum reconstruction defect $\delta(\Pi \circ \phi_t)$ grows under the coarse-graining projection Π from microstates to macrostates, where ϕ_t is the microstate dynamics. The second law is the statement that this growth is non-negative for measure-preserving ϕ_t.*

Corollary 8.3 (Equivalent Theories)**.** *Two physical theories are empirically indistinguishable at energy scale E if and only if $d_{\text{int}}(P_1^{(E)}, P_2^{(E)}) = 0$, where $P_i^{(E)}$ is the distribution of scattering outcomes at energy E predicted by theory i. The Derived Interface Conjecture asserts that Hilbert-space quantum gravity and Krein-space quadratic gravity satisfy this condition for all $E \ll M_{\text{Pl}}$.*

8.2. Cognition

Corollary 8.4 (Memory Loss)**.** *Cognitive memory loss is the growth of reconstruction defect $\Delta_R(h_{\text{past}})$ over time, where h_{past} is the generating history and R is the reconstruction map from current memory states to past states. Perfect recall requires injective memory encoding; all real memory systems have positive reconstruction defect.*

Corollary 8.5 (Learning as Defect Reduction)**.** *Learning is the reduction of task-relevant reconstruction defect: the update of the projection* Π *such that* $\delta(\Pi_{\text{new}}) < \delta(\Pi_{\text{old}})$ *for the task-relevant component of the defect. A learning event that reduces defect for one task while increasing it for another is a specialisation; one that reduces defect across all tasks is a genuine improvement.*

Corollary 8.6 (Paradigm Shift as Projection Change)**.** *A scientific or cognitive paradigm shift is a change of projection* $\Pi \to \Pi'$ *that is a distinction reorganisation:* Π' *is neither a refinement nor a coarsening of* Π*. The incommensurability of two paradigms is measured by* $\mathcal{D}(\Pi_*\mu, \Pi'_*\mu)$*: how different the observable interfaces they produce from the same underlying reality are.*

8.3. Institutions

Corollary 8.7 (Goodhart's Law)**.** *Goodhart's Law — when a measure becomes a target it ceases to be a good measure — is the theorem that optimising a proxy* M *for a true objective* T *navigates the fiber* $\Pi^{-1}(\textit{high } M)$ *in directions that may decrease* T*. The fiber is non-trivial whenever* Π *is non-injective, which it always is for any realistic measurement system. The severity of the Goodhart effect is measured by the distinction* $\mathcal{D}(P_M, P_T)$ *between the interface of the measure and the interface of the true objective.*

Corollary 8.8 (Institutional Opacity)**.** *An institution is opaque to degree* $\kappa > 0$ *at a given parameter if the sectional curvature of theory space in the direction of the institution's internal structure is* $+\kappa$*: many distinct internal configurations produce nearly identical observable behaviour. High positive curvature is high institutional opacity.*

Corollary 8.9 (Legibility as Injectivity)**.** *A policy or governance intervention is fully legible if the projection from the target community's internal dynamics to the policy's observable outcomes is injective. Scott's catastrophes arise when this projection is treated as injective (the community is treated as fully legible) when it is in fact highly non-injective (the community's internal dynamics are largely invisible to the policy measure).*

Conclusion

This paper has developed three interlocking mathematical structures and shown that they form a coherent account of what any observation does and what any comparison of observations reveals.

The reconstruction defect $\Delta_R(h) = d(h, R(\Pi(h)))$ measures the irreducible cost of non-injective projection: the minimum error any reconstruction must make, regardless of how cleverly it proceeds. The Irreversibility from Non-Injectivity theorem establishes that this cost is always positive when the projection is many-to-one, and the Entropy Bound connects the minimum defect to the Shannon entropy of the projected distribution. Irreversibility, memory loss, thermodynamic entropy, and the arrow of time are all instances of this same structure.

The distinction functional $\mathcal{D}(P, Q) = \mathrm{JSD}(P\|Q)$ measures the information destroyed when the distinction between two observable interfaces is forgotten. It is interface-native — defined without reference to any internal space — symmetric, and locally equivalent to the Fisher information metric. Gauge orbits are its zero-level sets. Hidden admissibility directions are deformations along which it remains zero. The Born rule of quantum mechanics is the simplest possible distinction functional: the map that forgets the phase of a complex amplitude.

The theory space $(\mathcal{M}, d_{\mathrm{int}})$ is the metric space of observable interfaces, equipped with the Jensen–Shannon distance. It is locally Riemannian with Fisher metric. Its curvature is the curvature of distinguishability: positive curvature means overfitting and opacity; negative curvature means high falsifiability. Geodesics are paths of minimal accumulated distinction. Theories at distance zero are empirically indistinguishable.

The unified translation of Theorem 5.1 collects all the major concepts of physical, cognitive, and institutional theory into a single framework: gauge orbits, ghost states, thermodynamic irreversibility, the Born rule, memory loss, Goodhart effects, and institutional opacity are all instances of the same three-level chain,

$$M \xrightarrow{\Pi} O \xrightarrow{\mathcal{D}} (\mathcal{M},\, d_{\mathrm{int}}),$$

differing only in what M, O, and Π are in each application.

The paper's argument can be compressed to a single substitution. The question

"What exists?"

is replaced by

"What distinctions survive projection, and with what geometry?"

The distinction functional and the theory space metric are the tools for answering the second question. They are the tools for doing science, epistemology, and institutional analysis without presupposing access to the internal space from which

observations are generated.

What remains is the investigation of specific instances. The monograph *Foundations of Distinction Geometry* develops the physics, cognitive, and institutional applications in detail. The present paper has established the mathematical core that those applications draw on.

References

[1] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory* **49**(7), 1858–1860 (2003).

[2] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, American Mathematical Society, Providence, 2000.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, Hoboken, 2006.

[4] N. Ay, J. Jost, H. Vân Lê, and L. Schwachhöfer, *Information Geometry*, Springer, Cham, 2017.

[5] I. Csiszár and P. Shields, "Information theory and statistics: a tutorial," *Foundations and Trends in Communications and Information Theory* **1**(4), 417–528 (2004).

[6] F. Nielsen, "On the Jensen–Shannon symmetrization of distances relying on abstract means," *Entropy* **21**(5), 485 (2019).

[7] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory* **37**(1), 145–151 (1991).

[8] D. Friedan, "Nonlinear models in $2 + \epsilon$ dimensions," *Physical Review Letters* **45**, 1057–1060 (1980).

[9] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, 2003.

[10] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference*, American Mathematical Society, Providence, 1982.