Kernel Embeddings and the Separation of Measure Phenomenon

Flyxion

June 2026

Based on Santoro, Waghmare & Panaretos, *PNAS* 123(23), 2026

Contents

Preface **ii**

1 The Two-Sample Problem and Its Difficulties **1**
1.1 The Classical Setting . . . 1
1.2 The Nonparametric Two-Sample Problem . . . 2
1.3 High-Dimensional and Complex Domains . . . 2
1.4 Kernel Methods and the MMD . . . 2
1.5 Overview of the Separation Phenomenon . . . 3

2 Reproducing Kernel Hilbert Spaces **4**
2.1 Kernels and Feature Maps . . . 4
2.2 Examples of Kernels . . . 5
2.3 Universality and Characteristic Kernels . . . 5
2.4 Operator Theory on Hilbert Spaces . . . 5
2.5 Mean and Covariance Embeddings . . . 6

3 Gaussian Measures in Infinite Dimensions **7**
3.1 Gaussian Measures on Hilbert Spaces . . . 7
3.2 Support of a Gaussian Measure . . . 8
3.3 Mutual Singularity and Equivalence . . . 9

4 The Feldman–Hájek Theorem **11**
4.1 Statement of the Theorem . . . 11
4.2 Geometric Interpretation . . . 12
4.3 The Criterion for Centered Gaussians . . . 12
4.4 Historical Notes . . . 13

5 The Main Theorem: Separation of Measure **14**
5.1 Setup and Assumptions . . . 14
5.2 The Separation of Measure Theorem . . . 14
5.3 Proof Strategy . . . 15
5.3.1 Step 1: Injectivity of the Covariance Embedding . . . 16
5.3.2 Step 2: The Special Structure of Kernel Covariance Operators 16
5.4 Why Covariance, Not Mean, Suffices . . . 17
5.5 A $0/\infty$ Law for the Likelihood Ratio . . . 18

6 The Blessing of Infinite Dimensionality **19**
6.1 Inverting the Curse . . . 19
6.2 Why Infinite Dimensionality Helps . . . 19
6.3 The Reformulation Principle . . . 20
6.4 Comparison with the Curse of Dimensionality . . . 20

6.5 A Historical Parallel: the Weierstrass Inversion 21

7 The Maximum Mean Discrepancy and Its Limitations 24
7.1 The MMD Revisited 24
7.2 Why MMD Does Not Exploit Separation 24
7.3 The Spectrum of Covariance Operators and Testing 25

8 Likelihood Ratio Tests by Kernel Gaussian Embedding 26
8.1 The Gaussian Likelihood Ratio 26
8.2 Regularisation 27
8.3 Theoretical Guarantees 27

9 Representation as Distinction Amplification 28
9.1 A Hierarchy of Representations 28
9.2 Distinguishability at Each Level 28
9.3 Formal Notion of Amplification 30
9.4 The Representation Chain 31
9.5 Representation-Induced Phase Transitions 31
9.6 Comparison: Projection-Induced Collapse 32

10 Information-Theoretic and Geometric Perspectives 33
10.1 The Information Geometry of the Result 33
10.2 Reachability and Support Geometry 33
10.3 Amplification vs. Preservation 34
10.4 Information Is Not Distinguishability 34
10.5 Distinguishability Phase Transitions: A General Theory 35
10.6 The Geometry-Distinguishability Meta-Principle 36
10.7 The Cameron–Martin Paradox as Conceptual Keystone 37

11 Admissibility, Reachability, and Constraint Geometry 38
11.1 An Interpretive Reframing: From Metric to Reachability 38
11.2 Strata as Admissible Regions (Interpretive) 39
11.3 Representational Choices and What They Preserve 39
11.4 Connections to CLIO and Projection-Induced Collapse 40
11.5 Semantic Geometry and Gaussian Measures (Speculative) 40

12 Extensions and Open Problems 41
12.1 Beyond Two-Sample Testing 41
12.2 Atomic Measures 41
12.3 Dependent Observations 42
12.4 Other Embeddings 42
12.5 Minimax Optimality 42

13 Summary and Conclusions 43
13.1 What Has Been Proved 43
13.2 Proof Architecture 43
13.3 Broader Significance 44
13.4 Closing Remarks 44

A Proof of the Feldman–Hájek Theorem 45
A.1 Reduction to the Diagonal Case 45

A.2 Kakutani's Product Theorem 45
A.3 Connecting to Hilbert–Schmidt 46

B Operator Norms and the Hilbert–Schmidt Space 47
B.1 The Hilbert–Schmidt Space 47
B.2 Trace-Class Operators and Their Role 47

C Background on Locally Compact Polish Spaces 48
C.1 Polish Spaces 48
C.2 Local Compactness 48

Preface

> "Simply put, we don't know what differences to look for, the possibilities are bewildering."
>
> Victor M. Panaretos (2026)

Two probability distributions walk into an infinite-dimensional Hilbert space. Under the right embedding, they will never meet again.

This is the informal content of the separation of measure theorem—with an important caveat: the result holds for *continuous* (non-atomic) distributions embedded via a *universal* kernel. Both conditions are essential. Discrete or atomic measures, and non-universal kernels, are outside the theorem's scope. The reader should keep this in mind whenever the exposition speaks loosely of "two distributions": throughout, these are non-atomic Borel probability measures and the kernel is assumed bounded, continuous, and universal.

This monograph is an extended exposition of a single mathematical theorem and its consequences: the *separation of measure phenomenon* identified by Santoro, Waghmare, and Panaretos in their 2026 *Proceedings of the National Academy of Sciences* paper. The theorem proves, in a remarkably clean and general form, that kernel covariance embeddings transform distinct continuous probability distributions into mutually singular Gaussian measures. Mutual singularity is the strongest possible form of separation: the two measures live on essentially disjoint regions of the ambient space, each invisible to the other.

The result matters for several reasons.

First, it provides a precise geometric explanation for an important aspect of the empirical success of kernel methods in high-dimensional two-sample testing. Practitioners have long observed that kernel-based tests outperform classical alternatives, but the mechanism was not well understood. The separation of measure phenomenon supplies a previously missing geometric explanation: suitable embedding places the hypotheses at infinite information-theoretic distance from one another at the population level, rendering the population-level distinction perfectly identifiable.

Second, the proof technique is beautiful. It routes through the classical Feldman–Hájek dichotomy, a deep theorem in the theory of Gaussian measures on infinite-dimensional Hilbert spaces that dates to the 1950s and 1960s. The paper shows that the two-sample testing problem can be *exactly* reformulated as the problem of distinguishing singular from equivalent Gaussian measures—a problem for which the Feldman–Hájek theorem provides a complete criterion. This is a paradigm example

of an old result in pure mathematics illuminating a contemporary question in data science.

Third, the paper demonstrates a negative companion result: mean embeddings alone do *not* produce separation. The covariance structure is essential. This is not merely a technical distinction; it reflects a deep fact about what kind of information is required to characterise a probability measure—a fact that resonates with longstanding themes in the philosophy of probability and information theory.

This monograph aims to make the paper accessible to a reader with a background in mathematics or theoretical statistics who may not be familiar with Gaussian measures on Hilbert spaces or the Feldman–Hájek theory. Accordingly, the first several chapters are devoted to foundations: reproducing kernel Hilbert spaces, operator theory on Hilbert spaces, Gaussian measures in infinite dimensions, and the Feldman–Hájek theorem itself. These are not mere preliminaries; each contains material that is worth understanding deeply for its own sake, and the main theorem can only be properly appreciated against this background.

Later chapters engage with the broader intellectual implications of the result: its connection to the information geometry of hypothesis testing, its relationship to the maximum mean discrepancy and other kernel criteria, the construction of practically implementable tests that exploit the separation phenomenon, and the philosophical question of what it means for an embedding to *amplify* rather than merely preserve distinctions.

A concluding chapter situates the result within several broader research programmes: process ontology and constraint geometry, semantic manifold theory, and the general question of how representational choices determine what distinctions are accessible to an observer.

The reader who desires only a route through the proof may follow Chapters **??**–4–5. The reader who desires to understand the result fully is encouraged to read in sequence.

June 2026 *Flyxion*

Chapter 1

The Two-Sample Problem and Its Difficulties

> "Without knowing which kinds of deviations to target, it becomes difficult to optimize the choice of test statistic."
>
> Santoro, Waghmare, Panaretos (2026)

1.1 The Classical Setting

The *two-sample problem* is one of the oldest and most fundamental questions in statistics: given observations $X_1, \dots, X_n \sim \mathbb{P}$ and $Y_1, \dots, Y_m \sim \mathbb{Q}$, does $\mathbb{P} = \mathbb{Q}$? This question arises wherever one wishes to detect differences between populations, compare experimental and control groups, evaluate generative models, or test whether two datasets could plausibly have the same source.

The formal version is a hypothesis test:

$$H_0 : \mathbb{P} = \mathbb{Q} \qquad \text{vs} \qquad H_1 : \mathbb{P} \neq \mathbb{Q}. \tag{1.1}$$

In low-dimensional settings, classical procedures are available. The *Kolmogorov–Smirnov test* compares empirical distribution functions; the *t-test* compares means under normality; the *Wilcoxon rank-sum test* is a nonparametric alternative robust to departures from normality. The chi-squared test handles categorical or discretized data.

The challenge, however, lies in what happens when the state space $\mathcal{X}$ is high-dimensional, structured, or of a type for which there is no natural total order—such as a space of functions, graphs, images, or text documents. The classical tests all rely, in some way, on the one-dimensional structure of the real line, whether through empirical CDFs, ranks, or scalar summaries. They do not transfer.

1.2 The Nonparametric Two-Sample Problem

The nonparametric version of the problem, where $\mathbb{P}$ and $\mathbb{Q}$ are arbitrary Borel probability measures on a general space $\mathcal{X}$ with no structural assumptions, is genuinely difficult. The alternative hypothesis H_1 is vast: there are infinitely many ways in which two distributions can differ. A difference might be confined to the mean while leaving higher moments untouched, or conversely present in the variance while the means coincide. It may reside in the tails rather than the bulk, in a multimodal structure invisible in the marginals, in the correlation structure without any marginal effect, or localized in a small region of the support rather than diffusely spread everywhere.

Any test statistic is, by definition, a summary of the data. A summary compresses information. The difficulty is that no scalar or finite-dimensional summary can adequately represent all possible forms of departure from H_0. A test that is powerful against mean shifts may have no power against higher-moment alternatives. A test designed for diffuse alternatives may miss sharp localized ones.

This is not merely a practical difficulty—it reflects a fundamental information-theoretic constraint. The nonparametric alternative is simply too large for any fixed choice of test statistic to be uniformly powerful.

1.3 High-Dimensional and Complex Domains

The difficulty is further compounded in modern applications. In genomics, each observation may be a gene expression profile with tens of thousands of dimensions. In natural language processing, each observation may be a document embedded in a high-dimensional latent space. In computer vision, each observation is a high-resolution image. In these settings, the state space $\mathcal{X}$ may be a subset of $\mathbb{R}^d$ for large d, a space of functions, a space of graphs or networks, a Riemannian manifold, a metric space with no algebraic structure, or a product of several heterogeneous spaces.

In each case, the structure of the space creates new possibilities for how $\mathbb{P}$ and $\mathbb{Q}$ can differ, and existing test statistics must be adapted or entirely redesigned.

1.4 Kernel Methods and the MMD

The most successful general-purpose approach to the nonparametric two-sample problem is the *kernel trick*. Instead of comparing $\mathbb{P}$ and $\mathbb{Q}$ directly, one maps both distributions into a reproducing kernel Hilbert space and compares them there. The primary measure of discrepancy is the *Maximum Mean Discrepancy*:

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; k) = \|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|_{\mathcal{H}}, \tag{1.2}$$

the RKHS distance between the kernel mean embeddings of the two distributions. Under a characteristic kernel, $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; k) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, so the MMD is a proper metric on the space of probability measures.

The empirical MMD can be computed from samples and its asymptotic distribution under the null is known, enabling the construction of tests with controlled type-I error. Extensive empirical evidence demonstrates that MMD-based tests outperform classical alternatives, particularly in complex and high-dimensional settings. Yet until recently, a transparent mathematical explanation for this superior performance was lacking.

The Central Question. Why do kernel methods perform so well? Is there a mathematical reason for the empirical success of the kernel trick in two-sample testing, beyond the fact that it embeds distributions into a rich space? The paper of Santoro, Waghmare, and Panaretos gives a precise and striking answer.

1.5 Overview of the Separation Phenomenon

The main result—to be developed in full in subsequent chapters—can be stated informally as follows:

> *If $\mathbb{P} \neq \mathbb{Q}$ are two distinct continuous (non-atomic) probability measures on a locally compact Polish space $\mathcal{X}$, then their kernel Gaussian embeddings $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ and $\mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})$ are mutually singular.*

Mutual singularity means the two Gaussian measures are concentrated on *disjoint* subsets of the RKHS. The embedded distributions do not merely differ—they live in separate regions of the infinite-dimensional space with no overlap whatsoever.

Equivalently:

$$\mathbb{P} \neq \mathbb{Q} \iff \mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \perp \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}}). \tag{1.3}$$

This reformulates the two-sample testing problem from asking whether two arbitrary distributions differ (the hard nonparametric problem) to asking whether two Gaussian measures are equivalent or singular (a structurally cleaner problem). The Feldman–Hájek theorem, reviewed in Chapter 4, gives a complete criterion for the latter. At the population level, the distinction is perfectly identifiable; the finite-sample problem of how to exploit this structure efficiently is addressed in Chapter 8.

This is the separation of measure phenomenon, and it is, as Panaretos describes it, a *blessing of infinite dimensionality*.

Chapter 2

Reproducing Kernel Hilbert Spaces

> "The key idea is to map probability measures to vectors or functions in a reproducing kernel Hilbert space, thereby enabling the application of linear or multivariate methods directly to distributions."
>
> Santoro, Waghmare, Panaretos (2026)

2.1 Kernels and Feature Maps

Definition 2.1. Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *positive semidefinite kernel* (or simply *kernel*) if for every $n \in \mathbb{N}$, every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geqslant 0. \tag{2.1}$$

A kernel is *symmetric* if $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$.

Associated to each kernel is a canonical *feature map* $\varphi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ defined by $\varphi(x) = k(x, \cdot) =: k_x$. The feature map sends each point $x \in \mathcal{X}$ to the function $y \mapsto k(x, y)$.

The key object is the Hilbert space generated by these feature maps.

Definition 2.2 (Reproducing Kernel Hilbert Space)**.** Let $\mathcal{X}$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel. The *reproducing kernel Hilbert space* (RKHS) associated to k, denoted $\mathcal{H} = \mathcal{H}(k)$, is the completion of the pre-Hilbert space

$$\mathcal{H}^0 = \operatorname{span}\{k_x : x \in \mathcal{X}\} \tag{2.2}$$

under the inner product determined by $\langle k_x, k_y \rangle_{\mathcal{H}} = k(x, y)$.

The defining property of an RKHS is the *reproducing property*:

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H}, x \in \mathcal{X}. \tag{2.3}$$

This says that the evaluation functional $f \mapsto f(x)$ is continuous (hence representable) and is reproduced by the kernel. It is a powerful regularity property: functions in $\mathcal{H}$ cannot oscillate wildly, since their values are controlled by their norm.

2.2 Examples of Kernels

Example 2.3 (Gaussian RBF kernel)**.** On $\mathcal{X} = \mathbb{R}^d$, the *Gaussian radial basis function kernel* is

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{2.4}$$

for bandwidth $\sigma > 0$. This is the most commonly used kernel in practice. Its RKHS consists of smooth functions, and it is universal on compact subsets of $\mathbb{R}^d$.

Example 2.4 (Matérn kernels)**.** The family of *Matérn kernels* on $\mathbb{R}^d$ parameterizes smoothness through a parameter $\nu > 0$. For $\nu = 1/2$, one recovers the Ornstein–Uhlenbeck kernel $k(x, y) = \exp(-\|x - y\|/\ell)$. For $\nu \to \infty$, one approaches the Gaussian RBF kernel.

Example 2.5 (Polynomial kernel)**.** On $\mathbb{R}^d$, the *polynomial kernel* of degree p is $k(x, y) = (\langle x, y\rangle + c)^p$. Its RKHS is a finite-dimensional space of polynomials of degree at most p.

2.3 Universality and Characteristic Kernels

Two properties of kernels are particularly relevant for statistical applications.

Definition 2.6 (Universal kernel)**.** A kernel k on a locally compact Hausdorff space $\mathcal{X}$ is $C_0(\mathcal{X})$*-universal* (or simply *universal*) if its RKHS $\mathcal{H}(k)$ is dense in $C_0(\mathcal{X})$, the space of continuous functions vanishing at infinity, under the uniform norm.

Definition 2.7 (Characteristic kernel)**.** A kernel k is *characteristic* for a class $\mathcal{F}$ of probability measures on $\mathcal{X}$ if the mean embedding map $\mathbb{P} \mapsto m_{\mathbb{P}}$ is injective on $\mathcal{F}$.

Universal kernels are characteristic. The Gaussian RBF and Matérn kernels are universal on $\mathbb{R}^d$. For a bounded universal kernel, the mean embedding uniquely identifies each probability measure.

2.4 Operator Theory on Hilbert Spaces

Since the covariance embedding maps distributions to *operators* rather than vectors, we need some operator theory.

Let $\mathcal{H}$ be a separable Hilbert space. For vectors $f, g \in \mathcal{H}$, their *tensor product* is the rank-one operator

$$(f \otimes g)\, u = \langle g, u\rangle_{\mathcal{H}}\, f, \qquad u \in \mathcal{H}. \tag{2.5}$$

Definition 2.8. An operator $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ is *self-adjoint* if $\mathbf{A} = \mathbf{A}^*$; *positive semidefinite* (written $\mathbf{A} \succeq 0$) if $\langle \mathbf{A}h, h\rangle_{\mathcal{H}} \geqslant 0$ for all h; *compact* if the image of the unit ball is precompact; *Hilbert–Schmidt* if $\|\mathbf{A}\|_{\mathrm{HS}}^2 := \mathrm{trace}(\mathbf{A}^*\mathbf{A}) < \infty$; and *trace-class* if $\|\mathbf{A}\|_{\mathrm{tr}} := \mathrm{trace}(\sqrt{\mathbf{A}^*\mathbf{A}}) < \infty$.

One always has $\|\mathbf{A}\|_{\mathrm{op}} \leqslant \|\mathbf{A}\|_{\mathrm{HS}} \leqslant \|\mathbf{A}\|_{\mathrm{tr}}$. Trace-class operators are compact; Hilbert–Schmidt operators are compact. For a positive semidefinite compact operator $\mathbf{A}$, the spectral theorem gives an orthonormal basis $\{e_j\}$ and non-negative eigenvalues $\{\lambda_j\}$ with $\mathbf{A}e_j = \lambda_j e_j$ and $\lambda_j \to 0$. Then $\mathrm{trace}(\mathbf{A}) = \sum_j \lambda_j$.

2.5 Mean and Covariance Embeddings

We now define the central objects of the paper.

Definition 2.9 (Kernel mean embedding)**.** Let k be a bounded kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. The *kernel mean embedding* of $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ is the unique element $\mathbf{m}_{\mathbb{P}} \in \mathcal{H}$ satisfying

$$\langle \mathbf{m}_{\mathbb{P}}, f\rangle_{\mathcal{H}} = \int_{\mathcal{X}} f(u)\, d\mathbb{P}(u) \qquad \text{for all } f \in \mathcal{H}. \tag{2.6}$$

Equivalently, $\mathbf{m}_{\mathbb{P}} = \int k_u \, d\mathbb{P}(u)$ (Bochner integral).

Definition 2.10 (Kernel covariance embedding)**.** The *(uncentred) kernel covariance embedding* of $\mathbb{P}$ is the operator $\mathbf{S}_{\mathbb{P}} : \mathcal{H} \to \mathcal{H}$ defined by

$$\langle f, \mathbf{S}_{\mathbb{P}} g\rangle_{\mathcal{H}} = \int_{\mathcal{X}} f(u)\, g(u)\, d\mathbb{P}(u) \qquad \text{for all } f, g \in \mathcal{H}, \tag{2.7}$$

or equivalently, $\mathbf{S}_{\mathbb{P}} = \int k_u \otimes k_u \, d\mathbb{P}(u)$.

The covariance embedding $\mathbf{S}_{\mathbb{P}}$ is self-adjoint, positive semidefinite, and trace-class. Its eigenvalues $\{\lambda_j(\mathbf{S}_{\mathbb{P}})\}$ are non-negative and satisfy $\sum_j \lambda_j(\mathbf{S}_{\mathbb{P}}) = \mathrm{trace}(\mathbf{S}_{\mathbb{P}}) = \int k(u,u)\, d\mathbb{P}(u) < \infty$.

Remark 2.11. When k is a bounded universal kernel on $\mathcal{X}$ and $\mathbb{P}$ has full support on $\mathcal{X}$, the operator $\mathbf{S}_{\mathbb{P}}$ is injective ($\ker(\mathbf{S}_{\mathbb{P}}) = \{0\}$). The argument is that $\langle f, \mathbf{S}_{\mathbb{P}} f\rangle = \int f(u)^2 \, d\mathbb{P}(u) = 0$ implies $f = 0$ $\mathbb{P}$-almost surely, and since $\mathbb{P}$ has full support and $f \in \mathcal{H}$ is continuous, this forces $f \equiv 0$. We include this fact for completeness; it is not used critically elsewhere, and readers should note that the exact relationship between universality, full support, and injectivity in full generality requires care (see [19]).

The pair $(\mathbf{m}_{\mathbb{P}}, \mathbf{S}_{\mathbb{P}})$ constitutes the combined embedding of $\mathbb{P}$. The Gaussian associated to this pair is the object of central interest.

Chapter 3

Gaussian Measures in Infinite Dimensions

> "A fundamental result in the theory of Gaussian measures states that Gaussians are either mutually equivalent or mutually singular, with no intermediate case."
>
> Santoro, Waghmare, Panaretos (2026)

3.1 Gaussian Measures on Hilbert Spaces

In finite dimensions, a Gaussian measure $\mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^d$ is characterized by its mean vector $\mu \in \mathbb{R}^d$ and positive semidefinite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The extension to infinite-dimensional Hilbert spaces requires care, because the natural analogue of the covariance matrix—an operator—must be trace-class for the Gaussian measure to be well-defined.

Definition 3.1. A Borel measure μ on a separable Hilbert space $\mathcal{H}$ is *Gaussian* if for every $f \in \mathcal{H}$, the random variable $\langle \cdot, f\rangle$ (under μ) is a real Gaussian. A Gaussian measure μ is determined by its *mean vector* $\mathbf{m} = \int u\, d\mu(u) \in \mathcal{H}$ and its *covariance operator* $\mathbf{C} = \int (u-\mathbf{m})\otimes(u-\mathbf{m})\, d\mu(u)$, where $\mathbf{C}$ is a positive semidefinite trace-class operator. We write $\mu = \mathcal{N}(\mathbf{m}, \mathbf{C})$.

The condition that $\mathbf{C}$ be trace-class is necessary for the Gaussian measure to be a proper probability measure on $\mathcal{H}$. If $\mathbf{C}$ is bounded but not trace-class, there is no Gaussian measure with that covariance in the Hilbert space sense (though one may exist on a larger Banach space).

Example 3.2. Consider the RKHS $\mathcal{H}$ associated to the Gaussian RBF kernel on $\mathbb{R}^d$, and let $\mathbb{P}$ be any absolutely continuous probability measure on $\mathbb{R}^d$. Then $\mathbf{S}_{\mathbb{P}}$ is a positive semidefinite trace-class operator on $\mathcal{H}$, and $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ is a well-defined Gaussian measure on $\mathcal{H}$.

3.2 Support of a Gaussian Measure

Definition 3.3. The *Cameron–Martin space* (or *reproducing kernel Hilbert space*) of a Gaussian measure $\mu = \mathcal{N}(\mathbf{m}, \mathbf{C})$ on $\mathcal{H}$ is the range of $\mathbf{C}^{1/2}$:

$$\mathcal{H}_\mu = \mathfrak{R}(\mathbf{C}^{1/2}) \subset \mathcal{H}. \tag{3.1}$$

The Cameron–Martin space determines the absolutely continuous translations of the Gaussian measure and governs much of its geometric behaviour, but it must be carefully distinguished from the *topological support*.

Definition 3.4. The *topological support* of a Borel measure μ on $\mathcal{H}$ is the smallest closed set of full μ-measure:

$$\operatorname{supp}(\mu) = \overline{\mathfrak{R}(\mathbf{C}^{1/2})} + \mathbf{m} \subset \mathcal{H}. \tag{3.2}$$

For $\mu = \mathcal{N}(\mathbf{m}, \mathbf{C})$, the topological support equals the closure of the Cameron–Martin space translated by the mean. The Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ is generally *dense* in $\operatorname{supp}(\mu)$ but is a proper subset of it whenever $\mathbf{C}$ is compact (since the range of a compact operator is not closed unless it is finite-dimensional). The Cameron–Martin space is therefore typically strictly smaller than the support.

Remark 3.5. This distinction matters. One should not say "the support of $\mathcal{N}(0, \mathbf{C})$ is $\mathfrak{R}(\mathbf{C}^{1/2})$." The correct statement is: the topological support is $\overline{\mathfrak{R}(\mathbf{C}^{1/2})}$, and the Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ is a dense subspace of the support on which absolutely continuous translations live. It is the Cameron–Martin space, not the topological support per se, that governs singularity via the Feldman–Hájek conditions.

When $\mathcal{H}$ is infinite-dimensional, the Cameron–Martin space presents a genuine measure-theoretic paradox: it has *zero measure* under the very Gaussian it governs. That is,

$$\mu\big(\mathfrak{R}(\mathbf{C}^{1/2})\big) = 0. \tag{3.3}$$

The Gaussian $\mu = \mathcal{N}(0, \mathbf{C})$ assigns full measure to the topological support $\overline{\mathfrak{R}(\mathbf{C}^{1/2})}$, yet typical samples drawn from μ lie *outside* the Cameron–Martin space—they belong to the support but not to its dense core. In other words, the structure that governs absolutely continuous translations and the Feldman–Hájek conditions is itself invisible to the measure; μ cannot "see" its own Cameron–Martin space in the sense that it assigns that space measure zero. This is one of the genuinely strange features of infinite-dimensional Gaussian geometry and is the geometric engine behind the Feldman–Hájek dichotomy.

A central geometric fact is:

Proposition 3.6. *If $\mathbf{C}$ is trace-class (necessarily the case for a Gaussian measure on an infinite-dimensional Hilbert space), then $\mathcal{N}(0, \mathbf{C})$ assigns measure zero to every bounded subset of $\mathcal{H}$, and in particular to any finite-dimensional subspace. The Gaussian is, in a precise sense, concentrated in the "infinite-dimensional" part of $\mathcal{H}$.*

This is an analogue of the fact that in $\mathbb{R}^2$, a degenerate Gaussian can be concentrated on a line. In infinite dimensions, every Gaussian with trace-class covariance is, in a sense, degenerate from the perspective of the full Hilbert space. See Figure 3.1 for a two-dimensional illustration.

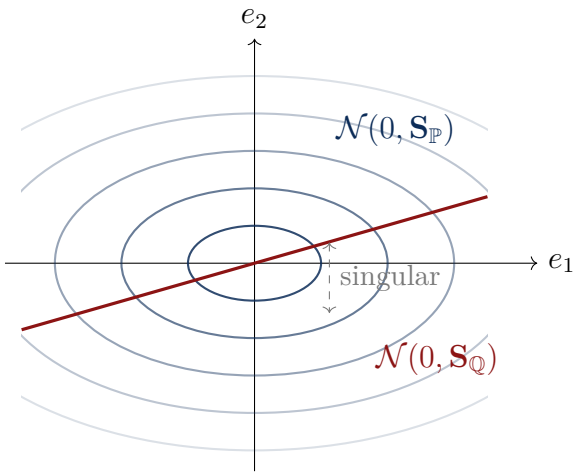

Figure 3.1: Illustration of mutual singularity of Gaussian measures in $\mathbb{R}^2$. One Gaussian is supported on the full plane (elliptical contours); the other is degenerate and supported on a line. The two measures are mutually singular: each is concentrated on a set of measure zero for the other. In infinite dimensions, the geometry is analogous but far richer.

3.3 Mutual Singularity and Equivalence

We now define the central dichotomy.

Definition 3.7. Two probability measures μ and ν on a measurable space $(\Omega, \mathcal{F})$ are *equivalent* (written $\mu \sim \nu$) if they have the same null sets, meaning $\mu(A) = 0 \iff \nu(A) = 0$ for all $A \in \mathcal{F}$. They are *mutually singular* (written $\mu \perp \nu$) if there exists a set $A \in \mathcal{F}$ with $\mu(A) = 1$ and $\nu(A) = 0$.

Mutual singularity is the strongest possible form of difference between measures: the Kullback–Leibler divergence is infinite in both directions, and neither measure has a density with respect to the other. Two mutually singular measures are concentrated on essentially disjoint sets—they are, in a precise sense, orthogonal.

Equivalence means they share the same null sets; each has density with respect to the other.

The remarkable fact, which lies at the heart of the Feldman–Hájek theorem, is that for Gaussian measures on Hilbert spaces, these are the *only* two possibilities.

Theorem 3.8 (Gaussian dichotomy)**.** *Two Gaussian measures on a separable Hilbert space are either equivalent or mutually singular. There is no intermediate case.*

This is in sharp contrast to the situation in finite dimensions. In $\mathbb{R}^d$, two Gaussian measures $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ with *nonsingular* covariance matrices are always equivalent: both are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, hence with respect to each other, and the Radon–Nikodym derivative is a smooth positive function. (The nondegeneracy assumption is essential: if either covariance matrix is singular, the corresponding Gaussian is supported on a proper affine subspace of $\mathbb{R}^d$, and two Gaussians supported on different affine subspaces are mutually singular even in finite dimensions—as the figure examples earlier

illustrate.) In infinite dimensions, even measures with nonsingular covariance operators can be mutually singular, because the Hilbert–Schmidt condition involves all infinitely many eigenvalues and may fail even when none of them is zero.

Chapter 4

The Feldman–Hájek Theorem

> "Our proof leverages the classical Feldman–Hájek dichotomy, and shows that even a small perturbation of a continuous distribution will be maximally magnified through its Gaussian embedding."
>
> Santoro, Waghmare, Panaretos (2026)

4.1 Statement of the Theorem

The Feldman–Hájek theorem gives an explicit criterion for when two Gaussian measures on a Hilbert space are equivalent versus singular. It is named for Jacob Feldman [5] and Jaroslav Hájek [6], who independently proved versions of it in 1958.

Theorem 4.1 (Feldman–Hájek)**.** *Let $\mu_i = \mathcal{N}(\mathbf{m}_i, \mathbf{C}_i)$, $i = 1, 2$, be two Gaussian measures on a separable Hilbert space $\mathcal{H}$. Then $\mu_1 \sim \mu_2$ or $\mu_1 \perp \mu_2$. The measures are equivalent if and only if the following three conditions hold simultaneously. First, the Cameron–Martin spaces coincide: $\mathfrak{R}(\mathbf{C}_1^{1/2}) = \mathfrak{R}(\mathbf{C}_2^{1/2})$. Second, the relative perturbation is Hilbert–Schmidt: $\mathbf{C}_1^{-1/2}\mathbf{C}_2\mathbf{C}_1^{-1/2} - \mathbf{I} \in \mathrm{HS}(\mathcal{H})$. Third, the mean shift lies in the Cameron–Martin space: $\mathbf{m}_1 - \mathbf{m}_2 \in \mathfrak{R}(\mathbf{C}_1^{1/2})$. If any one of these conditions fails, then $\mu_1 \perp \mu_2$.*

The theorem is sharp: failure of *any single one* of the three conditions implies mutual singularity.

Remark 4.2. Conditions 1 and 2 together say that the covariance operators must be *close* in a precise operator-theoretic sense. Condition 1 requires that the Cameron–Martin spaces of the two measures coincide—equivalently, that the two measures are supported on the same closed affine subspace (the closure of the Cameron–Martin space). Condition 2 says that the relative perturbation of the covariance operators, when expressed in a common eigenbasis, must be square-summable: it is not enough that the covariance operators differ by a trace-class perturbation; the perturbation must be Hilbert–Schmidt-close in a *relative* sense. Condition 3 says the mean shift must lie in the Cameron–Martin space $\mathfrak{R}(\mathbf{C}_1^{1/2})$, not merely in the topological support.

4.2 Geometric Interpretation

The Feldman–Hájek theorem has a clean geometric interpretation. Two Gaussians with different Cameron–Martin spaces have different absolutely-continuous translation classes: translations that preserve the measure for one will be singular for the other. If two Gaussians have different Cameron–Martin spaces—equivalently, if $\mathfrak{R}(\mathbf{C}_1^{1/2}) \neq \mathfrak{R}(\mathbf{C}_2^{1/2})$—they are mutually singular. If they have the same Cameron–Martin space but the eigenvalue ratios $\lambda_{2,j}/\lambda_{1,j}$ deviate from 1 in a non-square-summable way, they are also singular. Equivalence requires both conditions: matching Cameron–Martin spaces and a Hilbert–Schmidt-close relative perturbation.

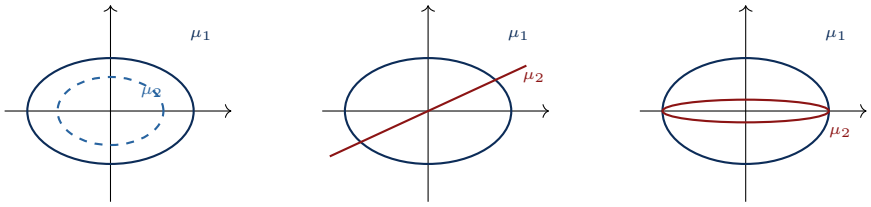

Figure 4.1: Schematic of the Feldman–Hájek dichotomy illustrated in $\mathbb{R}^2$. Left: Two Gaussian measures with the same support are equivalent (each has density with respect to the other). Centre: Two Gaussians with different supports (one full-plane, one on a line) are mutually singular. Right: A heuristic visualization only—in $\mathbb{R}^2$, two nondegenerate Gaussians are always equivalent regardless of how different their covariances are; the right panel is intended to suggest, schematically, the infinite-dimensional phenomenon where a highly anisotropic perturbation of the covariance causes failure of the Hilbert–Schmidt condition and hence mutual singularity. The actual phenomenon cannot be visualized in finite dimensions.

4.3 The Criterion for Centered Gaussians

For the purpose of the main theorem, the centered case—where both Gaussians have mean zero—is particularly important. In this case, Condition 3 of Theorem 4.1 is automatically satisfied (since $\mathbf{m}_1 = \mathbf{m}_2 = 0$), and the dichotomy simplifies:

Corollary 4.3. *For centered Gaussians $\mathcal{N}(0, \mathbf{C}_1)$ and $\mathcal{N}(0, \mathbf{C}_2)$:*

$$\mathcal{N}(0, \mathbf{C}_1) \sim \mathcal{N}(0, \mathbf{C}_2) \iff \begin{cases} \mathfrak{R}(\mathbf{C}_1^{1/2}) = \mathfrak{R}(\mathbf{C}_2^{1/2}) \\ \mathbf{C}_1^{-1/2}\mathbf{C}_2\mathbf{C}_1^{-1/2} - \mathbf{I} \in \mathrm{HS}(\mathcal{H}). \end{cases} \tag{4.1}$$

In all other cases, $\mathcal{N}(0, \mathbf{C}_1) \perp \mathcal{N}(0, \mathbf{C}_2)$.

Remark 4.4. The condition that $\mathfrak{R}(\mathbf{C}_1^{1/2}) = \mathfrak{R}(\mathbf{C}_2^{1/2})$ can be understood spectrally. If $\mathbf{C}_i$ has eigenvalues $\{\lambda_{i,j}\}_{j\geqslant 1}$ in the same eigenbasis, then the condition requires that the two covariance operators have the same null space, and the additional

Hilbert–Schmidt condition requires

$$\sum_{j=1}^{\infty} \left(\frac{\lambda_{2,j}}{\lambda_{1,j}} - 1 \right)^2 < \infty. \tag{4.2}$$

This is a fine-grained condition on how the eigenvalues must be comparable.

4.4 Historical Notes

The Feldman–Hájek theorem is one of the foundational results of the theory of Gaussian measures on infinite-dimensional spaces, developed in the 1950s–1970s alongside the theory of Gaussian processes and stochastic differential equations. Key contributions include Feldman [5], Hájek [6], and the comprehensive treatments of Skorokhod [8] and Bogachev [7]. The theorem has applications in statistics (testing problems for Gaussian processes), in quantum field theory (where functional integrals require understanding when measures are equivalent), and in optimal transport (since singularity has implications for transport costs).

The insight of Santoro, Waghmare, and Panaretos is to use the Feldman–Hájek theorem not as a tool for analyzing Gaussian processes directly, but as a lens through which to view *non-Gaussian* distributions via their Gaussian embeddings. This represents a novel application of a classical result.

Chapter 5

The Main Theorem: Separation of Measure

> "Even a small perturbation of a continuous distribution will be maximally magnified through its Gaussian embedding."
>
> Santoro, Waghmare, Panaretos (2026)

5.1 Setup and Assumptions

We now have all the ingredients needed to state and discuss the main result.

Let $\mathcal{X}$ be a locally compact Polish space. This is a broad class of spaces, encompassing $\mathbb{R}^d$ and open subsets thereof, compact metric spaces, countable discrete spaces such as $\mathbb{N}$ or $\mathbb{Z}^d$, Riemannian manifolds, and function spaces with appropriate topologies. The key conditions are local compactness and the Polish property (complete separable metric space). These ensure that measure theory works well and that the RKHS theory applies.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded, continuous, positive semidefinite kernel that is *universal* on $\mathcal{X}$. Recall that universality means $\mathcal{H}(k)$ is dense in $C_0(\mathcal{X})$. The Gaussian RBF kernel satisfies this condition on $\mathbb{R}^d$ and on any compact subset of $\mathbb{R}^d$.

A probability measure $\mathbb{P}$ on $\mathcal{X}$ is *non-atomic* (or continuous) if $\mathbb{P}(\{x\}) = 0$ for every $x \in \mathcal{X}$. This excludes point masses and atomic components.

5.2 The Separation of Measure Theorem

Theorem 5.1 (Separation of Measure; Santoro–Waghmare–Panaretos 2026)**.** *Let $\mathcal{X}$ be a locally compact Polish space, k a bounded, continuous, universal kernel on $\mathcal{X}$, and $\mathbb{P}, \mathbb{Q}$ non-atomic Borel probability measures on $\mathcal{X}$. Then:*

$$\mathbb{P} \neq \mathbb{Q} \iff \mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \perp \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}}), \tag{5.1}$$

where $\mathbf{S}_{\mathbb{P}} = \int k_u \otimes k_u \, d\mathbb{P}(u)$ *and* $\mathbf{S}_{\mathbb{Q}} = \int k_u \otimes k_u \, d\mathbb{Q}(u)$ *are the kernel covariance embeddings.*

The biconditional is the full force of the theorem. The direction $\mathbb{P} = \mathbb{Q} \Rightarrow \mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) = \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})$ (hence trivially equivalent) is immediate from the definition. The content is the reverse: $\mathbb{P} \neq \mathbb{Q} \Rightarrow$ the centered Gaussian embeddings are *mutually singular*.

Corollary 5.2 (Uncentered version)**.** *Under the same hypotheses:*

$$\mathbb{P} \neq \mathbb{Q} \iff \mathcal{N}(\mathbf{m}_{\mathbb{P}}, \mathbf{S}_{\mathbb{P}}) \perp \mathcal{N}(\mathbf{m}_{\mathbb{Q}}, \mathbf{S}_{\mathbb{Q}}). \tag{5.2}$$

Thus both centered and uncentered Gaussian embeddings achieve separation. The centered version is the main theorem; the uncentered is a corollary.

5.3 Proof Strategy

The proof proceeds by applying the Feldman–Hájek theorem (Theorem 4.1). To show $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \perp \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})$ when $\mathbb{P} \neq \mathbb{Q}$, it suffices (by Corollary 4.3) to show that at least one of the following must hold: either the Cameron–Martin spaces differ, $\mathfrak{R}(\mathbf{S}_{\mathbb{P}}^{1/2}) \neq \mathfrak{R}(\mathbf{S}_{\mathbb{Q}}^{1/2})$, or the relative perturbation $\mathbf{S}_{\mathbb{P}}^{-1/2}\mathbf{S}_{\mathbb{Q}}\mathbf{S}_{\mathbb{P}}^{-1/2} - \mathbf{I}$ fails to be Hilbert–Schmidt.

Caution 5.3. A common misreading of the theorem is to reason as follows: "$\mathbb{P} \neq \mathbb{Q}$ implies $\mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$ (by injectivity of the covariance embedding), and distinct trace-class operators on an infinite-dimensional Hilbert space *generically* fail the Feldman–Hájek conditions." This is *not* sufficient as a proof. It is easy to construct pairs of distinct trace-class operators for which equivalence still holds. For example, if

$$\lambda_j^{(2)} = \lambda_j^{(1)} \left(1 + \frac{1}{j^2}\right), \tag{5.3}$$

then $\mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$, yet

$$\sum_{j=1}^{\infty} \left(\frac{\lambda_j^{(2)}}{\lambda_j^{(1)}} - 1\right)^2 = \sum_{j=1}^{\infty} \frac{1}{j^4} < \infty, \tag{5.4}$$

so the Hilbert–Schmidt condition is satisfied and the Gaussians remain equivalent. The separation phenomenon is therefore *not* a consequence of generic operator-theoretic reasoning. It is a consequence of a *special structural property* of covariance operators that arise from non-atomic probability measures through kernel embedding.

The actual argument requires identifying what it is about *kernel-embedded* covariance operators that forces failure of the Feldman–Hájek conditions when $\mathbb{P} \neq \mathbb{Q}$. We outline this in two steps.

5.3.1 Step 1: Injectivity of the Covariance Embedding

The first step establishes that for a universal kernel, the covariance embedding is injective on non-atomic measures.

Proposition 5.4. *Let k be a bounded universal kernel on a locally compact Polish space $\mathcal{X}$. If $\mathbb{P}$ and $\mathbb{Q}$ are non-atomic probability measures with $\mathbf{S}_{\mathbb{P}} = \mathbf{S}_{\mathbb{Q}}$, then $\mathbb{P} = \mathbb{Q}$.*

Proof sketch. If $\mathbf{S}_{\mathbb{P}} = \mathbf{S}_{\mathbb{Q}}$, then for all $f, g \in \mathcal{H}$:

$$\int f(u)\, g(u)\, d\mathbb{P}(u) = \int f(u)\, g(u)\, d\mathbb{Q}(u). \tag{5.5}$$

This says that $\mathbb{P}$ and $\mathbb{Q}$ agree on all integrals of functions of the form $h = fg$ with $f, g \in \mathcal{H}$. The key step is to show that the linear span of such products is dense in $C_0(\mathcal{X})$ in the uniform norm.

This density claim is the substantive content of the argument and requires care. It does not follow directly from the definition of universality (which asserts density of $\mathcal{H}$ itself, not of products of elements of $\mathcal{H}$). The argument proceeds as follows: since k is universal, $\mathcal{H}$ is dense in $C_0(\mathcal{X})$. The pointwise products of elements of $\mathcal{H}$ belong to the RKHS $\mathcal{H}(k^2)$ of the squared kernel $k^2(x, y) = k(x, y)^2$. One then shows that $\mathcal{H}(k^2)$ is also dense in $C_0(\mathcal{X})$ using the Stone–Weierstrass theorem and the separating/non-vanishing properties of $\mathcal{H}$, provided k is a continuous universal kernel on a locally compact Polish space (see [19], Theorem 4.58 and surrounding discussion; and [2] for the version applicable here).

Given this density, $\mathbb{P}$ and $\mathbb{Q}$ agree on a class dense in $C_0(\mathcal{X})$, and by standard measure theory they agree on all bounded continuous functions, hence $\mathbb{P} = \mathbb{Q}$. □

Remark 5.5. The density of products of RKHS functions is a non-trivial fact that deserves explicit acknowledgment. In particular, injectivity of the covariance embedding is a *stronger* result than injectivity of the mean embedding (which only requires $\mathcal{H}$ to be dense, not products of elements of $\mathcal{H}$). The reader who wishes to see the full argument is directed to the original paper and the cited references.

This establishes: $\mathbb{P} \neq \mathbb{Q} \Rightarrow \mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$. But as the caution above shows, this is only the beginning.

5.3.2 Step 2: The Special Structure of Kernel Covariance Operators

The deep step is to show that the specific way in which $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ differ—arising as they do from non-atomic measures via the kernel covariance embedding—forces failure of the Feldman–Hájek Hilbert–Schmidt condition.

The key structural property is that $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ have the integral representations

$$\mathbf{S}_{\mathbb{P}} = \int_{\mathcal{X}} k_u \otimes k_u \, d\mathbb{P}(u), \qquad \mathbf{S}_{\mathbb{Q}} = \int_{\mathcal{X}} k_u \otimes k_u \, d\mathbb{Q}(u), \tag{5.6}$$

so their difference $\Delta = \mathbf{S}_{\mathbb{Q}} - \mathbf{S}_{\mathbb{P}}$ inherits the full measure-theoretic structure of the signed measure $\mathbb{Q} - \mathbb{P}$. When $\mathbb{P} \neq \mathbb{Q}$, this difference is non-zero and non-trivial throughout the RKHS.

The actual proof of why this forces failure of the Hilbert–Schmidt condition—that is, why

$$\sum_{j=1}^{\infty} \sigma_j \left(\mathbf{S}_{\mathbb{P}}^{-1/2} \, \Delta \, \mathbf{S}_{\mathbb{P}}^{-1/2} \right)^2 = +\infty \tag{5.7}$$

when $\mathbb{P} \neq \mathbb{Q}$ and both are non-atomic—requires detailed spectral arguments specific to covariance operators arising from kernel embeddings of continuous measures. These arguments are worked out in full in Santoro, Waghmare, and Panaretos [2].

Caution 5.6. It is tempting to explain the Hilbert–Schmidt failure by saying the difference "propagates across all eigenmodes of $\mathbf{S}_{\mathbb{P}}$." This is an intuition, not a proof. Universality of the kernel does not in itself imply rapid or slow spectral decay of $\mathbf{S}_{\mathbb{P}}$; spectral decay depends on kernel smoothness, the geometry of $\mathcal{X}$, and the regularity of $\mathbb{P}$. The theorem does not reduce to a statement about eigenvalue growth rates, and informal accounts that suggest it does are overstating what has been proved. The actual mechanism is subtler: it depends on structural properties of the integral representation of $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ that are incompatible with the Hilbert–Schmidt condition once both measures are non-atomic and distinct.

Remark 5.7. The non-atomicity condition is a genuine structural necessity. For atomic measures (finite mixtures of point masses), the covariance operators $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ have finite rank, the eigenspectrum terminates at a finite index N, and the Hilbert–Schmidt sum $\sum_{j=1}^{N}(\cdot)^2$ is automatically finite. The separation phenomenon is therefore intrinsically a property of *continuous* distributions.

The full proof is given in the original paper. The conceptual upshot is:

> **The Structural Miracle.** It is not merely that $\mathbb{P} \neq \mathbb{Q}$ implies $\mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$. The miracle is that the specific structure of kernel-embedded covariance operators arising from non-atomic measures forces the Feldman–Hájek Hilbert–Schmidt condition to fail whenever $\mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$. This is a non-trivial property of the kernel embedding that does not hold for arbitrary pairs of distinct trace-class operators.

5.4 Why Covariance, Not Mean, Suffices

A companion negative result shows that mean embeddings alone do not produce separation.

Proposition 5.8 (Mean embeddings do not separate; Santoro–Waghmare–Panaretos 2026)**.** *For any covariance operator $\mathbf{C}$ on $\mathcal{H}$ and any distinct non-atomic probability measures $\mathbb{P} \neq \mathbb{Q}$ on $\mathbb{R}$ whose mean embeddings satisfy $\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}} \in \mathfrak{R}(\mathbf{C}^{1/2})$, we have*

$$\mathcal{N}(\mathbf{m}_{\mathbb{P}}, \mathbf{C}) \sim \mathcal{N}(\mathbf{m}_{\mathbb{Q}}, \mathbf{C}). \tag{5.8}$$

More precisely: for any fixed covariance operator $\mathbf{C}$, there always exist distinct non-atomic measures $\mathbb{P} \neq \mathbb{Q}$ whose mean-shifted Gaussian embeddings are equivalent. The failure of mean-based separation is therefore not a rare edge case but a generic phenomenon.

The mechanism is Condition 3 of the Feldman–Hájek theorem: $\mathcal{N}(\mathbf{m}_1, \mathbf{C}) \sim \mathcal{N}(\mathbf{m}_2, \mathbf{C})$ if and only if $\mathbf{m}_1 - \mathbf{m}_2 \in \mathfrak{R}(\mathbf{C}^{1/2})$. Since $\mathfrak{R}(\mathbf{C}^{1/2})$ is a *dense* subspace of $\mathcal{H}$, the set of mean differences that preserve equivalence is dense in $\mathcal{H}$; for any target mean difference one can find measures achieving it. The condition is therefore generically satisfied, not exceptional.

This result is significant: it explains *why* the MMD, based on mean embeddings, is theoretically limited in its ability to exploit the separation phenomenon. The covariance structure is essential for achieving perfect separation.

Summary of the Main Result.

The covariance embedding $\mathbb{P} \mapsto \mathbf{S}_\mathbb{P}$ is injective on non-atomic measures: $\mathbb{P} \neq \mathbb{Q} \Rightarrow \mathbf{S}_\mathbb{P} \neq \mathbf{S}_\mathbb{Q}$. However, this injectivity alone does not imply singularity of the Gaussian embeddings. The additional content of the theorem is that the specific structure of kernel-embedded covariance operators—arising from non-atomic measures via an infinite-dimensional integral over the feature map—forces the Feldman–Hájek Hilbert–Schmidt condition to fail whenever $\mathbf{S}_\mathbb{P} \neq \mathbf{S}_\mathbb{Q}$. This is a genuine structural property of kernel embeddings, not a consequence of generic operator theory. The mean embedding alone does not achieve this: for any fixed covariance operator $\mathbf{C}$, the Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ is dense in $\mathcal{H}$, so the set of mean differences preserving Gaussian equivalence is itself dense, and failure of mean-based separation is generic rather than exceptional.

5.5 A $0/\infty$ Law for the Likelihood Ratio

The mutual singularity has a striking consequence for the likelihood ratio.

Theorem 5.9 ($0/\infty$ law; Santoro–Waghmare–Panaretos 2026)**.** *Under the conditions of Theorem 5.1, the KL divergence between the Gaussian embeddings satisfies* $\mathrm{KL}(\mathcal{N}(0, \mathbf{S}_\mathbb{P}) \| \mathcal{N}(0, \mathbf{S}_\mathbb{Q})) \in \{0, +\infty\}$*: it is zero under* H_0 *(when* $\mathbb{P} = \mathbb{Q}$*) and infinite under* H_1 *(when* $\mathbb{P} \neq \mathbb{Q}$*).*

This $0/\infty$ law says that the embedded hypotheses are not merely separated—they are *maximally* separated in an information-theoretic sense. The Gaussian likelihood ratio test is the natural test statistic; it vanishes under the null and diverges under the alternative.

In practice, one works with regularised versions of this likelihood ratio. Santoro and Panaretos [3] develop this into a complete testing framework with consistency guarantees and remarkable empirical power.

Chapter 6

The Blessing of Infinite Dimensionality

> "High-dimensional geometry can have concrete implications for modern data science."
>
> Victor M. Panaretos (2026)

6.1 Inverting the Curse

The *curse of dimensionality* is a phrase coined by Richard Bellman to describe the exponential growth of sample complexity with dimension. In statistical learning, it manifests as the observation that many methods require sample sizes that scale badly with the dimension of the input space. For nonparametric two-sample testing, the curse is severe: in dimension d, the minimax-optimal rate for testing at precision ε typically scales as $n \sim \varepsilon^{-(d+2)/d}$ or worse, depending on the smoothness class.

The separation of measure phenomenon represents an inversion of this intuition. Rather than high dimensions making the problem harder, embedding into an *infinite*-dimensional space makes the problem—in a limiting, idealized sense—trivially easy: distinct continuous distributions are perfectly separated. The blessing of infinite dimensionality is that the limiting geometry is simpler, not more complex.

6.2 Why Infinite Dimensionality Helps

The mechanism can be understood intuitively as follows.

In finite-dimensional Euclidean space $\mathbb{R}^d$, two Gaussian measures $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ with *nonsingular* covariance matrices $\Sigma_1 \neq \Sigma_2$ are always equivalent: both are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$ and hence with respect to each other. (Singular covariance matrices—degenerate Gaussians supported on lower-dimensional subspaces—can produce singularity even in finite dimensions, as already noted.) Distinguishing nondegenerate Gaussians in finite dimensions always requires statistical work.

In infinite-dimensional Hilbert space, the same Gaussians may be singular. The reason is that with infinitely many eigenvalues, even a small multiplicative perturbation $(\lambda_{2,j}/\lambda_{1,j} - 1)^2$ at each eigenmode can sum to infinity:

$$\sum_{j=1}^{\infty} \left(\frac{\lambda_{2,j}}{\lambda_{1,j}} - 1 \right)^2 = +\infty, \tag{6.1}$$

which by the Feldman–Hájek theorem implies mutual singularity. In finite dimensions, there are only finitely many terms and the sum is always finite.

Infinite dimensionality amplifies perturbations by accumulating their effects across all modes. A perturbation that would be invisible in any finite-dimensional projection becomes detectable through its cumulative effect across infinitely many modes.

6.3 The Reformulation Principle

The separation of measure theorem can be understood as a *reformulation principle*: it takes a hard problem (nonparametric two-sample testing) and reformulates it as an easier problem (distinguishing singular versus equivalent Gaussians).

This is a recurring strategy in mathematics. The real roots of a polynomial become easier to count by passing to $\mathbb{C}$, as the fundamental theorem of algebra attests. Linear ODEs become algebraic equations under Fourier or Laplace transform. Group-theoretic questions become linear-algebraic questions via representation theory, and topological questions become algebraic questions via cohomology. In each case, passing to a larger, more structured space reveals regularity that was hidden in the original setting.

The separation of measure theorem is an instance of the same principle applied to statistical testing. The RKHS is the richer space; the Gaussian embedding provides the reformulation; and the Feldman–Hájek theorem is the algebraic criterion that makes the reformulated problem tractable.

6.4 Comparison with the Curse of Dimensionality

The curse of dimensionality concerns the difficulty of estimation problems in high but *finite* dimensions. The blessing of infinite dimensionality concerns a qualitative phenomenon that emerges only in the limit. These two effects are not contradictory; they operate at different levels.

The practical takeaway is that although the idealized problem becomes perfectly separable in the RKHS, finite samples still require careful estimation of the relevant statistics. The blessing identifies what one should target; finite-sample methods determine how efficiently that target can be reached.

	Curse of Dimensionality	**Blessing of Infinite Dimensionality**
Domain	Finite $\mathbb{R}^d$, d large	Infinite-dimensional RKHS
Effect	Sample complexity grows exponentially	Distinct distributions are perfectly separated
Mechanism	More parameters to estimate	Eigenvalue sums diverge
Regime	Statistical / estimation	Information-theoretic / structural
Implication	Harder to distinguish distributions	Trivial to distinguish distributions (in principle)

Table 6.1: Comparison of the curse and blessing of dimensionality.

6.5 A Historical Parallel: the Weierstrass Inversion

The separation of measure phenomenon belongs to a recurring pattern in the history of analysis, and the pattern has a canonical first instance. Before the 1870s, the working intuition of mathematics held that continuity and smoothness were nearly the same property. A continuous curve was expected to be differentiable except at isolated singularities, because every curve one could draw looked locally straight under sufficient magnification. Weierstrass's function

$$f(x) \;=\; \sum_{n=0}^{\infty} a^n \cos\!\big(b^n \pi x\big), \qquad 0 < a < 1, \quad ab \geq 1, \tag{6.2}$$

destroyed this intuition with two ingredients in deliberate tension: the amplitude factor a^n decays fast enough to force uniform convergence, hence continuity, while the frequency factor b^n grows fast enough that fresh oscillations arrive at every scale. There is no magnification at which the graph becomes approximately straight. The derivative exists nowhere. The function is globally coherent but locally directionless — continuous in position, discontinuous in direction.

The lasting shock, however, was not the example but its typicality. Banach and Mazurkiewicz showed in 1931 that nowhere-differentiable functions are *residual* in $C[0,1]$: the smooth functions, far from being the norm, form a meagre set. The same verdict is returned by every inequivalent notion of "most" that has since been tried: prevalence in the measure-theoretic sense appropriate to spaces without translation-invariant measure, and almost-sure roughness of Brownian sample paths in the probabilistic sense. The monster was the generic object all along; the smooth curve was the measure-zero exception on which human geometric intuition happened to be trained.

The separation theorem runs the same inversion in the space of measures. Finite-dimensional probability trains the intuition that distinct distributions are merely

metrically separated: distinguishable at some finite distance, with mutual singularity a strong and special condition. Indeed, in finite dimensions any two nondegenerate Gaussians are equivalent, and this fact generalises silently into the expectation that distributions can always be partially distinguished but never perfectly separated. The theorem of Santoro, Waghmare, and Panaretos shows that after kernel covariance embedding this expectation fails in the strongest possible way: mutual singularity is the *generic* outcome for distinct non-atomic measures, and equivalence is the exceptional regime — exactly the regime that finite-dimensional experience samples. Once again the intuitive case is the measure-zero island, and the "pathological" case is the sea.

The parallel extends beyond the moral to the mechanism, and at two points it is structurally precise.

Scale-distributed discrepancy. The Weierstrass function fails to be differentiable not because its irregularity is large at any one scale but because it is present at every scale: the contributions $a^n \cos(b^n \pi x)$ never become negligible relative to the resolution at which one inspects the graph. The failure of equivalence in the separation theorem has the same anatomy. The covariance operators $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ may be close in operator norm, but equivalence requires the relative perturbation $\mathbf{S}_{\mathbb{P}}^{-1/2}(\mathbf{S}_{\mathbb{Q}} - \mathbf{S}_{\mathbb{P}})\mathbf{S}_{\mathbb{P}}^{-1/2}$ to be Hilbert–Schmidt, and the discrepancy injected by the embedding is not localised at any finite set of eigenmodes. In both cases the object is destroyed not by a singularity somewhere but by a distributed roughness everywhere: no truncation captures it, because the obstruction lives in the tail.

Dimensional interstice. The graph of the Weierstrass function has Hausdorff dimension $D = 2 + \log_b a$, strictly between 1 and 2: too irregular to be a curve, not irregular enough to fill a region. The Feldman–Hájek stratification places a Gaussian measure in an analogous interstice. The Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ — the directions along which translation preserves equivalence — has measure zero under the Gaussian itself, while the measure is supported inside the full space $\mathcal{H}$. The Gaussian lives *between* its Cameron–Martin space and $\mathcal{H}$: not concentrated on the former, not spread over the latter, occupying a stratum that no integer-graded intuition anticipates. The fractional dimension of the Weierstrass graph and the measure-zero-yet-generating role of the Cameron–Martin space are two appearances of the same phenomenon: in infinite-codimension settings, the natural objects sit in graded interstices that finite-dimensional experience does not equip us to expect.

The common structure, stated once for both results: a construction that looks as if it should fail — the function ought to be differentiable somewhere; the embedded measures ought to be equivalent — instead succeeds in the maximally strong sense, and the success is not exceptional but generic. Weierstrass's function was the first warning that the space of possibilities is dominated by objects our intuition classifies as monsters, with the familiar objects forming a negligible exceptional set. Fractals, turbulence, Brownian motion, random graphs, and high-dimensional geometry each rediscovered the warning in their own vocabulary. The separation of measure phenomenon is the same inversion arriving in the theory of kernel embeddings: what looked like the strongest and rarest relation between distributions — living on disjoint regions of the space, each invisible to the other — turns out to be what distinct distributions *generically do*, the moment the ambient space is large enough to let

them.

A reader who knows either result therefore has an immediate foothold on the other. The Weierstrass function is the separation theorem of $C[0,1]$: it separates continuity from smoothness with a gap that is generic rather than constructed. The separation theorem is the Weierstrass function of $\mathcal{G}(\mathcal{H})$: it reveals that the equivalence classes our intuition treats as the ordinary case are meagre strata in a space whose generic relation is mutual singularity. Both are blessings in the same currency — the very roughness and immensity of infinite-dimensional spaces is what makes perfect separation, and hence perfect distinguishability, available at all.

Chapter 7

The Maximum Mean Discrepancy and Its Limitations

7.1 The MMD Revisited

The Maximum Mean Discrepancy is defined as

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k) = \|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|_{\mathcal{H}}^2 = \iint k(x, y)\, d(\mathbb{P} - \mathbb{Q})(x)\, d(\mathbb{P} - \mathbb{Q})(y). \tag{7.1}$$

Under a characteristic kernel, $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$. The empirical estimator

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2} \sum_{i,j} k(X_i, X_j) + \frac{1}{m^2} \sum_{i,j} k(Y_i, Y_j) - \frac{2}{nm} \sum_{i,j} k(X_i, Y_j) \tag{7.2}$$

is an unbiased estimator of MMD^2 and converges at the parametric $n^{-1/2}$ rate under the null.

7.2 Why MMD Does Not Exploit Separation

The key limitation of the MMD is identified by Proposition 5.8: it is based on the *mean* embedding and does not use the covariance. The mean embedding does not achieve separation of measure. Two distributions can differ substantially in their second-order structure while having identical mean embeddings; the MMD would then have zero power.

More formally: the test statistic $\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|_{\mathcal{H}}$ measures the distance between two *vectors* in $\mathcal{H}$. The separation of measure phenomenon, by contrast, lives in the structure of the covariance *operators*. These are fundamentally different objects—vectors versus operators—and the information in the operator cannot be recovered from the vector.

Remark 7.1. This does not mean the MMD is a poor test; in practice, it performs well because typical alternatives differ both in mean and covariance structure. The point is that it is not *designed* to exploit the separation phenomenon, and it can miss alternatives that manifest only at the covariance level.

7.3 The Spectrum of Covariance Operators and Testing

Since the separation of measure phenomenon is entirely a covariance effect, the natural test statistics should be functions of $\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}$ or of the relative spectrum of $\mathbf{S}_{\mathbb{Q}}$ with respect to $\mathbf{S}_{\mathbb{P}}$. Candidate statistics include the Hilbert–Schmidt distance $\|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|_{\mathrm{HS}}^2$ between covariance operators, the Gaussian KL divergence $\mathrm{KL}(\mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \| \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}}))$, and the regularised Gaussian likelihood ratio. The follow-up paper of Santoro and Panaretos [3] develops the likelihood-ratio approach, showing it achieves the $0/\infty$ law of Theorem 5.9 and yields tests with power exceeding state-of-the-art methods.

Chapter 8

Likelihood Ratio Tests by Kernel Gaussian Embedding

> "The likelihood ratio is specifically tailored to detect equality versus singularity of two Gaussians, and satisfies a $0/\infty$ law."
>
> Santoro & Panaretos (2025)

8.1 The Gaussian Likelihood Ratio

The Kullback–Leibler divergence between two Gaussian measures on a Hilbert space is only well-defined when the measures are equivalent; otherwise it is $+\infty$ by convention. In the equivalent case, for $\mathcal{N}(0, \mathbf{C}_1)$ and $\mathcal{N}(0, \mathbf{C}_2)$, it takes the form

$$\mathrm{KL}(\mathcal{N}(0, \mathbf{C}_1) \| \mathcal{N}(0, \mathbf{C}_2)) = \frac{1}{2}\Big[\mathrm{trace}(\mathbf{C}_2^{-1}\mathbf{C}_1 - \mathbf{I}) - \log\det(\mathbf{C}_2^{-1}\mathbf{C}_1)\Big], \tag{8.1}$$

where $\log\det$ denotes the operator log-determinant. In infinite dimensions, ordinary determinants do not exist for general bounded operators, so $\log\det$ requires a regularised interpretation.

Definition 8.1 (Fredholm log-determinant)**.** For a positive operator $\mathbf{A} = \mathbf{I} + \mathbf{T}$ where $\mathbf{T}$ is trace-class, the *Fredholm determinant* is

$$\det_1(\mathbf{A}) = \prod_{j=1}^{\infty} \lambda_j(\mathbf{A}), \tag{8.2}$$

and the *Fredholm log-determinant* is $\log\det(\mathbf{A}) = \log\det_1(\mathbf{A}) = \sum_{j=1}^{\infty} \log \lambda_j(\mathbf{A})$.

The Fredholm log-determinant is well-defined and finite precisely when $\mathbf{A} - \mathbf{I}$ is trace-class. Under the Feldman–Hájek equivalence conditions, the relative operator $\mathbf{C}_2^{-1}\mathbf{C}_1 - \mathbf{I}$ is Hilbert–Schmidt; if additionally $\mathrm{trace}(\mathbf{C}_2^{-1}\mathbf{C}_1 - \mathbf{I}) < \infty$, the Fredholm determinant applies and (8.1) is finite. The more general *Carleman–Fredholm determinant* $\det_2(\mathbf{A}) = \det_1(\mathbf{A}\, e^{-(\mathbf{A}-\mathbf{I})})$ is finite for Hilbert–Schmidt perturbations and is used in the regularised setting.

For the purposes of the separation theorem, the precise finiteness conditions matter less than the qualitative consequence. Under H_0 (when $\mathbb{P} = \mathbb{Q}$), we have $\mathbf{C}_1 = \mathbf{C}_2$, so $\mathbf{C}_2^{-1}\mathbf{C}_1 = \mathbf{I}$, giving $\operatorname{trace}(\mathbf{0}) = 0$ and $\log\det(\mathbf{I}) = 0$, hence KL $= 0$. Under H_1 (when $\mathbb{P} \neq \mathbb{Q}$ and both are non-atomic), the Gaussians are singular by Theorem 5.1, the Feldman–Hájek Hilbert–Schmidt condition fails, and KL $= +\infty$. This is the $0/\infty$ law of Theorem 5.9.

8.2 Regularisation

Since the measures are singular under the alternative, the raw likelihood ratio is $+\infty$ and cannot be used directly. The solution is regularisation: replace $\mathbf{S}_{\mathbb{Q}}$ with $\mathbf{S}_{\mathbb{Q}} + \alpha\mathbf{I}$ for a regularisation parameter $\alpha > 0$. This smooths the operators sufficiently to ensure the KL is finite, at the cost of some power.

The regularised test statistic is:

$$T_\alpha = \mathrm{KL}\Big(\mathcal{N}(0, \mathbf{S}_{\mathbb{P}} + \alpha\mathbf{I}) \,\Big\|\, \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}} + \alpha\mathbf{I})\Big), \tag{8.3}$$

estimated from data as $\hat{T}_\alpha$ using empirical covariance operators. The null distribution is obtained by permutation.

8.3 Theoretical Guarantees

Santoro and Panaretos [3] prove three main results for this test. First, *consistency*: as $n \to \infty$ and $\alpha \to 0$ at an appropriate rate, $\hat{T}_\alpha \to +\infty$ under H_1. Second, *uniform power*: the test achieves non-trivial power against all alternatives satisfying mild regularity conditions. Third, *unification*: the framework subsumes and extends spectral regularisation approaches based on the MMD.

Empirically, the likelihood-ratio test reports power substantially higher than MMD-based tests across a range of alternatives, confirming that exploiting the covariance structure yields real gains.

Chapter 9

Representation as Distinction Amplification

> "Suitable kernel embedding into an infinite-dimensional RKHS separates continuous distributions perfectly, even when their differences are arbitrarily subtle."
>
> Santoro, Waghmare, Panaretos (2026)

9.1 A Hierarchy of Representations

The separation of measure theorem is most naturally understood as a theorem about *representation*. The same probability measure $\mathbb{P}$ can be represented in three increasingly rich ways:

$$\text{Mean embedding:} \quad \mathbb{P} \mapsto \mathbf{m}_{\mathbb{P}} \in \mathcal{H}, \tag{9.1}$$

$$\text{Covariance embedding:} \quad \mathbb{P} \mapsto \mathbf{S}_{\mathbb{P}} \in \mathrm{TC}(\mathcal{H}), \tag{9.2}$$

$$\text{Gaussian embedding:} \quad \mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \in \mathcal{G}(\mathcal{H}), \tag{9.3}$$

where $\mathrm{TC}(\mathcal{H})$ denotes the space of positive trace-class operators on $\mathcal{H}$ and $\mathcal{G}(\mathcal{H})$ denotes the space of centered Gaussian measures on $\mathcal{H}$.

Each step in this hierarchy is strictly richer: the mean embedding is a special case of the Gaussian embedding (via the first moment), and the covariance captures second-order structure invisible to the mean alone. But the richness is not merely additive—it is qualitative. The Gaussian embedding lives in a space that carries an entirely different geometry: not the linear geometry of a Hilbert space, but the *information geometry* of probability measures, with its singularity/equivalence dichotomy.

9.2 Distinguishability at Each Level

We can precisely characterise what each representation is capable of distinguishing.

Proposition 9.1 (Distinguishability of the mean embedding)**.** *Let k be a characteristic kernel. Then:*

$$\mathbf{m}_{\mathbb{P}} = \mathbf{m}_{\mathbb{Q}} \iff \mathbb{P} = \mathbb{Q}. \tag{9.4}$$

The mean embedding is injective: it distinguishes all distributions. However, the distance $\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|_{\mathcal{H}}$ can be arbitrarily small even when $\mathbb{P} \neq \mathbb{Q}$, so the mean embedding preserves *distinctions but does not amplify them.*

Proposition 9.2 (Distinguishability of the covariance embedding)**.** *Let k be a bounded universal kernel and $\mathbb{P}, \mathbb{Q}$ non-atomic. Then:*

$$\mathbf{S}_{\mathbb{P}} = \mathbf{S}_{\mathbb{Q}} \iff \mathbb{P} = \mathbb{Q}. \tag{9.5}$$

The covariance embedding is also injective. As with the mean, the operator norm $\|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|_{\mathrm{op}}$ can be small for $\mathbb{P} \neq \mathbb{Q}$, so the covariance embedding alone also merely preserves rather than amplifies. The amplification occurs in the next step.

Theorem 9.3 (Distinguishability of the Gaussian embedding)**.** *Under the same hypotheses:*

$$\mathbb{P} \neq \mathbb{Q} \implies \mathrm{KL}(\mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \| \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})) = +\infty. \tag{9.6}$$

The Gaussian embedding not only distinguishes $\mathbb{P}$ from $\mathbb{Q}$: it places them at infinite *information-theoretic distance. This is amplification, not merely preservation.*

Theorem 9.3 is precisely the separation of measure phenomenon restated in amplification language. The passage from the covariance embedding to the Gaussian embedding is where the amplification occurs: the operator $\mathbf{S}_{\mathbb{P}}$ and the measure $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ contain the same data, but the measure-theoretic lens reveals that data at a categorically different scale.

Theorem 9.4 (Injectivity does not imply amplification)**.** *An injective representation $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{M}$ need not be amplifying. The mean embedding $\mathbb{P} \mapsto \mathbf{m}_{\mathbb{P}}$ is injective under a characteristic kernel, but for any $\varepsilon > 0$ there exist $\mathbb{P} \neq \mathbb{Q}$ with $\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|_{\mathcal{H}} < \varepsilon$: distinguishability under the mean embedding is finite and can be made arbitrarily small. The covariance embedding $\mathbb{P} \mapsto \mathbf{S}_{\mathbb{P}}$ is likewise injective under a universal kernel for non-atomic measures, but $\|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|_{\mathrm{HS}}$ can similarly be made arbitrarily small for $\mathbb{P} \neq \mathbb{Q}$. The Gaussian embedding $\mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$, by contrast, is amplifying: for any $\mathbb{P} \neq \mathbb{Q}$ (non-atomic), $\mathrm{KL}(\mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \| \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})) = +\infty$, so no matter how close $\mathbb{P}$ and $\mathbb{Q}$ are in any metric on $\mathcal{P}(\mathcal{X})$, their Gaussian embeddings are infinitely far apart.*

This theorem makes the key conceptual point precise. The covariance operator $\mathbf{S}_{\mathbb{P}}$ and the Gaussian measure $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ carry *identical information* about $\mathbb{P}$—the same data, encoded in operator eigenvalues. No new information is added when passing from $\mathbf{S}_{\mathbb{P}}$ to $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$. Yet the geometry of distinguishability undergoes a qualitative phase transition: from the metric space $(\mathrm{TC}(\mathcal{H}), \|\cdot\|_{\mathrm{HS}})$ with finite distances, to the information-geometric space of Gaussian measures where distinct non-atomic distributions are at infinite distance.

This is the deepest conceptual novelty of the paper: *the same information can simultaneously support finite distinguishability in one geometric setting and infinite distinguishability in another.* The choice of representation determines which geometric regime is accessible.

Proposition 9.5 (Information preservation under geometric reinterpretation)**.** *The map $\mathbf{S}_{\mathbb{P}} \mapsto \mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ does not add informational content. Specifically, the covariance operator $\mathbf{S}_{\mathbb{P}}$ and the centered Gaussian measure $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ are in bijective correspondence: given one, the other is uniquely determined. No data about $\mathbb{P}$ that is absent from $\mathbf{S}_{\mathbb{P}}$ is present in $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$.*

Nevertheless, the map strictly enlarges the class of admissible notions of distinguishability. At the level of operators, distinguishability is measured by operator norms ($\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_{\mathrm{HS}}$, $\|\cdot\|_{\mathrm{tr}}$), all of which produce finite real values. At the level of Gaussian measures, the natural divergence is the KL divergence under equivalence, or the singularity/equivalence dichotomy under the Feldman–Hájek stratification—a $\{0, +\infty\}$-valued structure. The latter notion is not available at the operator level.

The proposition states precisely what the theorem's philosophical heart is: not a gain in information, but a change in the geometry through which existing information is expressed. Moving from $\mathbf{S}_{\mathbb{P}}$ to $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ is a change of representation that unlocks a new and more powerful notion of distinguishability—one unavailable at the operator level—without adding any new content.

9.3 Formal Notion of Amplification

We can formalise the distinction between preservation and amplification using the language of *distinguishability functions*.

Definition 9.6. For a representation $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{M}$ (where $\mathcal{M}$ is a metric or information-geometric space with divergence D), the *distinguishability* of the pair $(\mathbb{P}, \mathbb{Q})$ under Φ is:

$$\mathrm{Dist}_{\Phi}(\mathbb{P}, \mathbb{Q}) = D(\Phi(\mathbb{P}), \Phi(\mathbb{Q})). \tag{9.7}$$

Under this definition, the mean distinguishability is $\mathrm{Dist}_{\mathrm{mean}}(\mathbb{P}, \mathbb{Q}) = \|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\|^2_{\mathcal{H}} = \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \in [0, \infty)$; the covariance distinguishability is $\mathrm{Dist}_{\mathrm{cov}}(\mathbb{P}, \mathbb{Q}) = \|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|^2_{\mathrm{HS}} \in [0, \infty)$ when finite; and the Gaussian distinguishability is $\mathrm{Dist}_{\mathrm{Gauss}}(\mathbb{P}, \mathbb{Q}) = \mathrm{KL}(\mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \| \mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})) \in \{0, +\infty\}$.

The Gaussian distinguishability collapses to a singular/equivalent dichotomy: either zero (when $\mathbb{P} = \mathbb{Q}$, measures are equivalent) or infinite (when $\mathbb{P} \neq \mathbb{Q}$, non-atomic, measures are singular). It does not produce a graded measure of similarity. Note that KL is not a metric and is not symmetric; the $\{0, +\infty\}$ structure here reflects the Feldman–Hájek dichotomy applied to the embedded Gaussians, not a distance in the usual sense. This is the mathematical content of the "blessing."

Remark 9.7. The singular/equivalent dichotomy of the Gaussian distinguishability—0 or ∞—is both a strength and a limitation. As a population-level theoretical property, it implies that any difference between non-atomic distributions is in principle detectable via kernel methods. As a practical guide, it indicates that the right test statistic should target this singularity structure (via the regularised likelihood ratio), rather than targeting the graded MMD distance. Finite-sample behaviour reintroduces grading through regularisation.

9.4 The Representation Chain

We can display the hierarchy of representations as a chain:

$$\mathbb{P} \xrightarrow{\text{mean}} \mathbf{m}_{\mathbb{P}} \xrightarrow{\text{covariance}} \mathbf{S}_{\mathbb{P}} \xrightarrow{\text{Gaussian}} \mathcal{N}(0, \mathbf{S}_{\mathbb{P}}) \xrightarrow{\text{FH dichotomy}} \{\text{strata}\}. \tag{9.8}$$

At each arrow, information is *reorganised*: no information is lost (all maps are injective), but the geometry in which information is expressed changes. The final step—from Gaussian measure to Feldman–Hájek stratum—is where the binary amplification occurs. Two distributions are in the same stratum if and only if they are equal.

The chain can be extended further (to higher-order moment embeddings, conditional embeddings, etc.), but the Gaussian step is the one that achieves the amplification from finite distinguishability to infinite distinguishability in a single move.

9.5 Representation-Induced Phase Transitions

Theorem 9.4 exhibits a qualitative phase transition in distinguishability topology:

$$\underbrace{\mathbb{P} \to \mathbf{m}_{\mathbb{P}}}_{\text{finite dist.}} \longrightarrow \underbrace{\mathbb{P} \to \mathbf{S}_{\mathbb{P}}}_{\text{finite dist.}} \longrightarrow \underbrace{\mathbb{P} \to \mathcal{N}(0, \mathbf{S}_{\mathbb{P}})}_{\text{infinite dist. } (\{0,\infty\})}. \tag{9.9}$$

We can characterise the type of representation by its effect on distinguishability:

Definition 9.8 (Classification of representations by distinguishability type)**.** Let $\Phi : \mathcal{P}(\mathcal{X}) \to (\mathcal{M}, D)$ be a representation. It is *collapsing* if there exist $\mathbb{P} \neq \mathbb{Q}$ with $D(\Phi(\mathbb{P}), \Phi(\mathbb{Q})) = 0$. It is *preserving* (or *metrically faithful*) if $D(\Phi(\mathbb{P}), \Phi(\mathbb{Q})) > 0$ whenever $\mathbb{P} \neq \mathbb{Q}$ while $D(\Phi(\mathbb{P}_n), \Phi(\mathbb{Q})) \to 0$ remains possible as $\mathbb{P}_n \to \mathbb{Q}$. It is *amplifying* if $D(\Phi(\mathbb{P}), \Phi(\mathbb{Q})) = +\infty$ whenever $\mathbb{P} \neq \mathbb{Q}$ within the relevant class.

Under this classification, projections and non-injective summaries are collapsing. The mean embedding (under a characteristic kernel) and the covariance embedding (under a universal kernel) are preserving. The Gaussian embedding, mapping covariance operators to Gaussian measures with KL divergence as D, is amplifying on non-atomic measures.

The kernel theorem is therefore a precise instance of a broader phenomenon: *representation-induced phase transitions in distinguishability topology*. The transition from preserving to amplifying is not a smooth increase in distinguishability; it is a jump from a graded real-valued metric to a binary $\{0, +\infty\}$ information structure. No intermediate stage exists, because of the Gaussian dichotomy theorem (Theorem 3.8).

This observation connects naturally to a range of other topics where representation choice determines what distinctions are accessible. In the theory of sufficient statistics, a sufficient statistic $T(X)$ is neither collapsing nor amplifying in the relevant sense; it preserves the likelihood ratio exactly. In information geometry, the Fisher–Rao metric on parametric families captures smooth, preserving geometry, while the kernel Gaussian embedding achieves discontinuous, amplifying geometry.

In the study of categorical versus continuous cognition, within-category variation is graded (preserving geometry) while between-category distinctions may be amplified to near-categorical separation. The kernel theorem is, in this light, a precise mathematical instance of the general principle that *information is not merely stored in representations—representations determine which distinctions become geometrically visible and at what scale.*

9.6 Comparison: Projection-Induced Collapse

To sharpen the concept of amplification, it helps to contrast it with the opposite phenomenon: *projection-induced collapse.*

A projection $\pi : \mathcal{X} \to \mathcal{Y}$ (for example, a dimension-reduction map, a discretisation, or a summary statistic) can map distinct states $x \neq x'$ to the same image $\pi(x) = \pi(x')$. Information is irreversibly lost. In measure-theoretic terms, if $\pi_*\mathbb{P} = \pi_*\mathbb{Q}$ for the pushforward measures, the projection cannot distinguish $\mathbb{P}$ from $\mathbb{Q}$ regardless of how much data is collected.

The kernel covariance embedding is the dual of this: an *injective* map from measures to a richer space that not only prevents collapse but actively amplifies distinctions. The amplification is maximal: no information about the relative position of $\mathbb{P}$ and $\mathbb{Q}$ is lost, and the representation reveals them as infinitely separated.

	Projection	**Kernel Gaussian Embedding**
Type	Many-to-one	Injective
Effect on distinctions	Collapses (may erase)	Amplifies to singularity
Distinguishability	$\leqslant$ original	Maximal (0 or ∞)
Dimension	Lower	Infinite
Information	Lost	Reorganised, amplified

Table 9.1: Projection versus embedding as dual operations on the space of probability measures.

Chapter 10

Information-Theoretic and Geometric Perspectives

10.1 The Information Geometry of the Result

The separation of measure phenomenon can be viewed through the lens of information geometry. The space of Gaussian measures on a Hilbert space is itself a metric space under the Fisher information metric (for finite-dimensional Gaussians, this gives the Fisher–Rao metric). The KL divergence is the natural information-theoretic divergence.

The $0/\infty$ law of Theorem 5.9 says that after Gaussian embedding, the hypotheses H_0 and H_1 are at zero distance and infinite distance respectively. There is no intermediate case. The embedded hypothesis space has been *collapsed* to a single point (the null) surrounded by an *abyss* (the alternatives at infinite information-theoretic distance).

This is an extreme form of information amplification. The MMD amplifies differences to $O(1)$ scale with $O(n^{-1/2})$ fluctuations, while the Gaussian likelihood ratio amplifies differences to $O(\infty)$ scale—they are infinite in the population—leaving finite-sample fluctuations as the operative constraint.

10.2 Reachability and Support Geometry

A natural reinterpretation of the separation theorem uses the language of *reachability*. Two probability measures are mutually singular when one cannot "reach" the other by any absolutely continuous transformation: there is no path of likelihood ratios connecting them.

For a Gaussian measure $\mathcal{N}(0, \mathbf{C})$, its Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ is the set of admissible shift directions—the directions $h \in \mathcal{H}$ for which the translated measure $\mathcal{N}(h, \mathbf{C})$ is absolutely continuous with respect to $\mathcal{N}(0, \mathbf{C})$. Points outside the closure $\overline{\mathfrak{R}(\mathbf{C}^{1/2})}$ (the topological support) are not reachable at all; typical sample points lie in the topological support but outside the Cameron–Martin space. If two Gaussians have different Cameron–Martin spaces, their absolutely-continuous translation classes are incompatible: no shift admissible for one is admissible for the

other, and the measures are mutually singular.

From this perspective: the kernel covariance embedding sends probability measures to operators, and operators to Cameron–Martin spaces (via the square-root map). Two distinct non-atomic measures are sent to operators whose Cameron–Martin spaces are geometrically incompatible—lying in different Feldman–Hájek strata. The singularity of the Gaussian measures reflects this incompatibility.

10.3 Amplification vs. Preservation

A theme in the paper is the distinction between *preserving* distinctions versus *amplifying* them. It is important to locate the amplification precisely.

The mean embedding $\mathbb{P} \mapsto \mathbf{m}_{\mathbb{P}}$ is injective under a characteristic kernel: it preserves the distinction between measures, but preserves them at the same scale. Two nearby measures have nearby mean embeddings, and two far-apart measures have far-apart mean embeddings. The distinction is preserved but not amplified.

The covariance embedding $\mathbb{P} \mapsto \mathbf{S}_{\mathbb{P}}$ is also injective and also merely preserving: the operator norm $\|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|_{\text{op}}$ or Hilbert–Schmidt norm $\|\mathbf{S}_{\mathbb{P}} - \mathbf{S}_{\mathbb{Q}}\|_{\text{HS}}$ can be made arbitrarily small for distinct non-atomic measures. At the level of operators, no amplification has occurred.

The amplification occurs in the final step: passing from $\mathbf{S}_{\mathbb{P}}$ to the Gaussian measure $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ and evaluating distinguishability through the Feldman–Hájek stratification. The covariance operator and the Gaussian measure encode the same information about $\mathbb{P}$; no new data is added. Yet the geometry of distinguishability changes categorically. An infinitesimally small difference between $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ in any operator norm becomes infinite KL divergence in the information-geometric sense. This is why the terminology must be precise: the *covariance embedding* is preserving; the *Gaussian embedding* is amplifying.

The analogy is to a change of coordinates that reveals structure already present but geometrically inaccessible: an isometry preserves distances, a projection collapses them, and a representation change can reveal distinctions that existed but were hidden in the original geometry. The kernel Gaussian embedding is such a representation change: it does not create information, it reorganises the geometry through which information is expressed.

10.4 Information Is Not Distinguishability

The deepest observation in this monograph can be stated as a single principle:

The Fundamental Asymmetry. Information and distinguishability are not the same thing. The covariance operator $\mathbf{S}_{\mathbb{P}}$ and the Gaussian measure $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ carry identical information about $\mathbb{P}$. Yet their distinguishability structures are categorically different: finite operator distances in the first case, infinite KL divergence in the second. Distinguishability is not a property of information itself. It is a property of the geometry in which information is

interpreted.

To make this precise, we introduce the following definition.

Definition 10.1 (Distinguishability structure)**.** Let $\mathcal{I}$ be a class of probability measures, and let $\Phi : \mathcal{I} \to (\mathcal{M}, D)$ be a representation into a space $\mathcal{M}$ equipped with a divergence D. The *distinguishability structure* induced by Φ is the function

$$d_\Phi : \mathcal{I} \times \mathcal{I} \to [0, +\infty], \qquad d_\Phi(\mathbb{P}, \mathbb{Q}) = D(\Phi(\mathbb{P}), \Phi(\mathbb{Q})). \tag{10.1}$$

Two representations Φ_1 and Φ_2 are *distinguishability-equivalent* if $d_{\Phi_1} = d_{\Phi_2}$ as functions on $\mathcal{I} \times \mathcal{I}$.

The key observation is that two representations can be related by a bijection—contain identical information—while inducing categorically different distinguishability structures. The passage $\mathbf{S}_\mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_\mathbb{P})$ is precisely such a bijection: injective in both directions, yet changing d_Φ from a finite-valued operator norm to a $\{0, +\infty\}$-valued information-theoretic structure. The practical utility of a representation is therefore not fully determined by the information it contains, but by the distinguishability structure it exposes.

10.5 Distinguishability Phase Transitions: A General Theory

The separation of measure theorem is a specific instance of a more general phenomenon. A representation $\Phi : \mathcal{I} \to (\mathcal{M}, D)$ induces a *distinguishability phase transition* if it is a bijection on $\mathcal{I}$ yet the distinguishability structure d_Φ is qualitatively different from the original—specifically, if the original structure has finite values everywhere while d_Φ takes the value $+\infty$ for all $\mathbb{P} \neq \mathbb{Q}$ within the relevant class.

The separation theorem is exactly this: $\Phi_1 = (\mathbb{P} \mapsto \mathbf{S}_\mathbb{P})$ gives finite operator distances, $\Phi_2 = (\mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_\mathbb{P}))$ gives $\{0, +\infty\}$-valued distinguishability, and both are injections on non-atomic measures. The bijection does not create information. It changes the ambient geometry, and the new geometry makes distinctions infinitely sharp.

This pattern recurs across mathematics in contexts that appear unrelated. In Fourier analysis, a function $f \in L^2(\mathbb{R})$ and its Fourier transform $\hat{f}$ carry identical information (Parseval's theorem ensures the map is an L^2-isometry), yet functions that overlap heavily in the time domain may be sharply separated in the frequency domain. The Fourier transform does not amplify information; it reorganises the geometry of function space so that certain distinctions become transparent.

In representation theory, a group element $g \in G$ and its matrix image $\rho(g)$ under a linear representation carry the same group-theoretic information, but in the representation space, algebraic relations become eigenvalue and trace conditions that are far easier to detect. The representation theory machine converts hard group-theoretic distinctions into accessible linear-algebraic ones.

In the theory of sufficient statistics, a sufficient statistic $T(X)$ for a parametric family $\{P_\theta\}$ carries the same information about θ as the full data X, yet the distinguishability structure for θ simplifies: the relevant comparison is now between the distributions of $T(X)$ under different θ, which may be much simpler geometrically than comparing the full data distributions.

In renormalization group theory in physics, a physical system at one scale is mapped to an equivalent description at another scale. The two descriptions contain the same information (near a renormalization fixed point), yet the distinguishability of different phases of matter—ordered versus disordered, gapped versus gapless—becomes visible only at the correct scale. The phase transition in the physical system manifests as a distinguishability phase transition in the representation.

In manifold learning, a high-dimensional data set embedded in $\mathbb{R}^d$ may contain geometric structure (cluster membership, intrinsic dimensionality, topological features) that is invisible in Euclidean distance but becomes apparent after nonlinear embedding. Algorithms such as UMAP and t-SNE seek representations in which the distinguishability structure of interest is geometrically accessible.

What these examples share with the kernel embedding theorem is the following structure: an information-preserving bijection $\Phi : \mathcal{I} \to \mathcal{M}$ such that d_Φ has qualitatively different properties from d_{id}. The bijection does not create information; it changes the ambient geometry, and the new geometry supports a richer or more accessible distinguishability.

10.6 The Geometry-Distinguishability Meta-Principle

The examples above suggest the following principle:

> **The Geometry-Distinguishability Meta-Principle.** The practical power to distinguish elements of a class $\mathcal{I}$ depends not only on the information they contain but on the geometry of the representation in which they are expressed. Information-preserving bijections can change the distinguishability structure without changing information content. The relevant question for a statistical or inferential problem is therefore not only which representation preserves the most information, but which representation exposes the relevant distinctions in the most accessible geometry.

This principle is not new to mathematicians. It is implicit in the use of Fourier analysis, representation theory, and sufficient statistics throughout modern mathematics. What is novel in the Santoro–Waghmare–Panaretos theorem is the precision with which it instantiates the principle: the precise class of objects (non-atomic Borel measures), the precise representation (kernel Gaussian embedding), the precise geometry (Gaussian measures with Feldman–Hájek stratification), and the precise form of the phase transition ($\{0, +\infty\}$-valued KL divergence versus finite operator distances).

The meta-principle also raises a natural research question. Given a class $\mathcal{I}$ and a statistical task, which representation Φ exposes the relevant distinctions most effectively? The separation theorem answers this for the two-sample testing problem with

kernel covariance Gaussian embeddings. For other tasks and other representation families, the analogous question is largely open. The general problem of designing representations that induce distinguishability phase transitions for specific inferential goals could be called *representational amplification design* and constitutes a research programme suggested by but extending beyond the paper.

10.7 The Cameron–Martin Paradox as Conceptual Keystone

No account of the representation-geometry relationship in the Gaussian setting is complete without returning to the Cameron–Martin paradox. Recall: for an infinite-dimensional Gaussian $\mu = \mathcal{N}(0, \mathbf{C})$,

$$\mu\big(\mathfrak{R}(\mathbf{C}^{1/2})\big) = 0. \tag{10.2}$$

The measure assigns zero probability to its own Cameron–Martin space.

This fact is the geometric engine behind the entire singularity phenomenon. In finite dimensions, the Cameron–Martin space of a nondegenerate Gaussian on $\mathbb{R}^d$ is all of $\mathbb{R}^d$, and the measure is everywhere positive. Every direction is admissible, every shift preserves absolute continuity, and two Gaussians with the same covariance matrix are always equivalent regardless of their means. The Cameron–Martin structure is invisible because it fills the whole space.

In infinite dimensions, the Cameron–Martin space $\mathfrak{R}(\mathbf{C}^{1/2})$ is a proper dense subspace of the topological support $\overline{\mathfrak{R}(\mathbf{C}^{1/2})}$, and the measure assigns it probability zero. Typical samples live outside the Cameron–Martin space: they belong to the support but not to its dense core. The admissible shift directions form a null set under the very measure they govern.

This creates a situation with no finite-dimensional analogue. The structure governing whether two Gaussians can reach one another (Cameron–Martin spaces, Hilbert–Schmidt conditions) is structurally present but empirically invisible. The Feldman–Hájek conditions are conditions on a null set from the measures' own perspective.

The philosophical implication is deep: in infinite dimensions, the *structural* properties of a measure and its *empirical* properties—what typical samples look like—are decoupled in a way that cannot occur in finite dimensions. Deciding whether two Gaussians are equivalent or singular is a purely structural question, unanswerable by any finite collection of samples. It requires analytic conditions on the operators.

This is why the separation theorem is simultaneously so powerful and so subtle. Its power is real and exact at the population level; the challenge of exploiting it in practice is exactly the challenge of estimating structural properties from empirical ones, which is what regularised likelihood-ratio tests address. The Cameron–Martin paradox—the measure cannot see its own Cameron–Martin space—is not a pathology to be explained away. It is the source of the theorem's force, and understanding it is understanding why infinite-dimensional Gaussian geometry is fundamentally different from anything finite-dimensional intuition can anticipate.

Chapter 11

Admissibility, Reachability, and Constraint Geometry

> "The important effect is information amplification. A good representation does not merely retain distinctions. It can magnify them until they become geometrically unavoidable."
>
> Commentary on the paper

Note on scope. The preceding chapters have been mathematical exposition: definitions, theorems, and proof sketches closely following Santoro, Waghmare, and Panaretos [2]. This chapter is different in character. It offers an *interpretation* of the theorem's meaning, reframing the separation of measure result within a broader conceptual vocabulary of reachability, admissible regions, and representational hierarchy. Nothing in this chapter is claimed to be a consequence of the paper; these are readings of what the theorem *means*, not additional mathematical results derived from it. The distinction between exposition and interpretation is kept explicit throughout.

11.1 An Interpretive Reframing: From Metric to Reachability

One way to read the separation of measure theorem is as a shift from metric questions to reachability questions. Classical statistical distances—total variation, Hellinger, KL divergence—are metrics on the space of probability measures, producing non-negative real numbers. The question they answer is: *how far apart* are $\mathbb{P}$ and $\mathbb{Q}$?

The separation of measure theorem, as interpreted here, asks instead: do $\mathbb{P}$ and $\mathbb{Q}$ *occupy the same region* of the ambient space after embedding? The answer is binary (equivalent or singular). This is a *reachability* question: two measures are mutually reachable (equivalent) if and only if they assign positive probability to

exactly the same events. If any event has positive probability under one but zero under the other, they are unreachable from one another (singular).

This reachability framing is an interpretation, not a theorem. The paper does not use the language of reachability. But the mathematical content—that $\mathbb{P} \neq \mathbb{Q}$ places the embedded Gaussians in disjoint support strata with no path of absolutely continuous transformations between them—makes this interpretation natural.

11.2 Strata as Admissible Regions (Interpretive)

The Feldman–Hájek theorem partitions the space of Gaussian measures on $\mathcal{H}$ into equivalence classes under the relation of mutual absolute continuity. Two Gaussians are in the same class ("stratum") if and only if they are equivalent—they have the same Cameron–Martin space and Hilbert–Schmidt-close covariances. This stratification is a mathematical fact.

The *interpretive* step is to call these strata "admissible regions" and to say that the kernel covariance embedding "sends each non-atomic measure to its own stratum." This language of admissibility is not standard in the paper; it is borrowed from the vocabulary of process ontology and constraint geometry to illuminate the structure. The mathematical content it glosses is precise: the separation theorem says the embedding $\mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ is injective at the level of Feldman–Hájek strata, mapping distinct non-atomic measures to distinct equivalence classes.

11.3 Representational Choices and What They Preserve

The distinction between mean embeddings and covariance embeddings is an instance of a general principle: *representational choices determine what distinctions are accessible*.

The mean embedding $\mathbb{P} \mapsto \mathbf{m}_{\mathbb{P}}$ represents each distribution by a single vector in $\mathcal{H}$. Vectors in $\mathcal{H}$ form a linear space; the geometry is flat. Distinctions are preserved but not amplified.

The covariance embedding $\mathbb{P} \mapsto \mathbf{S}_{\mathbb{P}}$ represents each distribution by a trace-class operator. Operators in $\mathcal{H}$ form a nonlinear space with richer structure. The Feldman–Hájek theorem is sensitive to operator structure in ways that have no finite-dimensional analogue.

The Gaussian embedding $\mathbb{P} \mapsto \mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ represents each distribution by a probability measure on $\mathcal{H}$. The space of probability measures carries the information geometry (Fisher metric, KL divergence), which is where singularity and equivalence live.

Each successive representation is richer, and each successive richness enables one to see distinctions that were invisible at the previous level. This is a hierarchy of representations, each amplifying distinctions left unresolved by the previous.

11.4 Connections to CLIO and Projection-Induced Collapse

In certain theoretical frameworks, one encounters the phenomenon of *projection-induced collapse*: different underlying states are mapped to the same representation by a many-to-one projection. Information is lost.

The separation of measure phenomenon demonstrates the opposite: an embedding that *expands* distinctions until they are maximally separated.

A projection $\pi : \mathcal{X} \to \mathcal{Y}$ is many-to-one and typically finite-dimensional; distinctions may be erased entirely. The kernel Gaussian embedding $\mathcal{P}(\mathcal{X}) \to \mathcal{G}(\mathcal{H})$, by contrast, is injective and infinite-dimensional; distinctions are amplified to singularity. These are dual operations, and the study of which distinctions are preserved, erased, or amplified by a given representation is a central question in the theory of information and inference.

11.5 Semantic Geometry and Gaussian Measures (Speculative)

The following extrapolation goes beyond anything claimed or implied in the original paper and is offered as a direction for future work.

If documents, concepts, or memories are modelled as probability measures over a latent feature space, the kernel covariance embedding would map each semantic object to a Gaussian measure on an RKHS. If the conditions of the separation theorem hold—the measures are non-atomic and the kernel is universal—then distinct semantic objects would be mapped to mutually singular Gaussian measures. They would be perfectly distinguishable in the embedding space.

The semantic geometry of the space would then be determined by the structure of the Gaussian measures: two semantic objects are "close" if their Gaussian embeddings are equivalent (same Cameron–Martin space, Hilbert–Schmidt-close covariances) and "far apart" if the embeddings are singular (different strata). This gives a qualitative two-level geometry in which intra-stratum distances are continuous and smooth (captured by the Fisher metric), while inter-stratum distinctions are topologically disconnected (captured by singularity). The intra-stratum geometry captures graded similarity; the inter-stratum structure captures categorical difference, echoing the distinction between within-category and between-category variation in conceptual spaces.

Chapter 12

Extensions and Open Problems

12.1 Beyond Two-Sample Testing

The separation of measure theorem has the potential to inform inference tasks beyond two-sample testing. Several directions are natural:

Classification

If each class is modelled by a probability distribution $\mathbb{P}_c$, the Gaussian embeddings $\mathcal{N}(0, S_{\mathbb{P}_c})$ are mutually singular across classes. A Bayes classifier in the RKHS would achieve zero Bayes error in the population limit. Finite-sample classifiers exploiting this structure could yield improved performance.

Generative Model Evaluation

Training a generative model and evaluating whether the generated distribution $\hat{\mathbb{P}}$ equals the target $\mathbb{P}$ is a two-sample testing problem. The separation theorem implies that even subtle mode-dropping or artefact injection would be detectable in principle. The likelihood-ratio test provides a tool for this.

Variational Inference

Variational inference minimizes the KL divergence between a variational distribution q and a target p. If q is parameterized as a Gaussian on an RKHS, the Feldman–Hájek conditions provide a criterion for when the approximation is exact versus when it remains in a different stratum.

12.2 Atomic Measures

The theorem requires non-atomicity. For atomic measures (including empirical measures), the covariance operator has finite rank, and the Feldman–Hájek conditions can be satisfied even for distinct measures. This is not a deficiency—empirical measures are always distinct from each other—but it highlights that the theorem is fundamentally a population-level result.

The finite-sample theory requires regularisation precisely because empirical measures are atomic. The regularisation parameter interpolates between the atomic and non-atomic regimes.

12.3 Dependent Observations

The theorem as stated assumes independent samples from $\mathbb{P}$ and $\mathbb{Q}$. Extensions to time series data with temporal dependence, spatial data with spatial autocorrelation, and samples from a Markov chain after mixing would require modifications to both the embedding and the test statistic. The separation phenomenon itself may hold under weaker conditions, but the precise conditions and the form of the test statistic would need to be worked out.

12.4 Other Embeddings

The paper considers the kernel covariance embedding, but other natural embeddings are available: higher-order moment embeddings $\mathbb{P} \mapsto \int k_x^{\otimes p} \, d\mathbb{P}(x)$ for $p \geqslant 3$, conditional embeddings for conditional distributions, and distributional embeddings into other function spaces such as Sobolev or Besov spaces. Whether analogues of the separation theorem hold for these embeddings is open.

12.5 Minimax Optimality

The likelihood-ratio test of Santoro and Panaretos achieves consistency and uniform power under mild conditions. A natural question is whether it achieves minimax-optimal power rates against specific classes of alternatives—for instance, against alternatives at a given smoothness level in the RKHS. Establishing minimax lower bounds and matching upper bounds would complete the statistical theory.

Chapter 13

Summary and Conclusions

13.1 What Has Been Proved

The main contribution of Santoro, Waghmare, and Panaretos is a clean, general, and mathematically rigorous theorem:

> **Theorem (informal).** Let $\mathbb{P}$ and $\mathbb{Q}$ be distinct continuous probability distributions on a locally compact Polish space. Let k be a bounded universal kernel with RKHS $\mathcal{H}$, and let $\mathbf{S}_{\mathbb{P}}$, $\mathbf{S}_{\mathbb{Q}}$ be the corresponding kernel covariance embeddings. Then the Gaussian measures $\mathcal{N}(0, \mathbf{S}_{\mathbb{P}})$ and $\mathcal{N}(0, \mathbf{S}_{\mathbb{Q}})$ are mutually singular.

This theorem achieves five things. It provides a precise geometric explanation for an important aspect of the empirical success of kernel methods in two-sample testing. It reformulates the nonparametric two-sample problem as a Gaussian singularity detection problem. It identifies the covariance structure, rather than the mean, as the essential ingredient. It yields a $0/\infty$ law for the likelihood ratio, pointing toward a new generation of powerful tests. And it demonstrates a "blessing of infinite dimensionality" that inverts the naive intuition that high dimensions make problems harder.

13.2 Proof Architecture

The proof routes through the classical Feldman–Hájek theorem in two non-trivial steps. The first establishes injectivity of the kernel covariance embedding on non-atomic measures (Proposition 5.4), giving $\mathbb{P} \neq \mathbb{Q} \Rightarrow \mathbf{S}_{\mathbb{P}} \neq \mathbf{S}_{\mathbb{Q}}$. The second, and deeper, step is structural: the specific way covariance operators $\mathbf{S}_{\mathbb{P}}$ and $\mathbf{S}_{\mathbb{Q}}$ arise from non-atomic measures via the kernel integral representation forces the Feldman–Hájek Hilbert–Schmidt condition to fail. This does not follow from generic operator theory, since distinct trace-class operators need not produce singular Gaussians. The conclusion then follows immediately from the Feldman–Hájek theorem: since the Hilbert–Schmidt condition fails, the Gaussian embeddings are mutually singular. The elegance of this proof lies in its reuse of old machinery (Feldman–Hájek, 1958)

in a new context (kernel embeddings, 2000s–2020s), combined with the structural insight that kernel covariance operators carry information across the entire spectrum in a way incompatible with Gaussian equivalence.

13.3 Broader Significance

Beyond the specific result, the paper illustrates several broader principles:

Representation determines accessibility. The choice of how to represent a probability measure—as a mean embedding (vector), a covariance embedding (operator), or a Gaussian measure—determines what information is accessible and at what scale. Richer representations enable finer distinctions.

Amplification is possible. Good representations do not merely preserve distinctions; they can amplify them. The kernel covariance embedding amplifies even infinitesimal differences between distributions to infinite information-theoretic distance.

Old mathematics illuminates new problems. The Feldman–Hájek theorem was proved in 1958, long before the modern theory of kernel methods. Its application to the two-sample testing problem is a reminder that classical analysis often contains tools that await novel applications.

The population limit reveals structure. The separation phenomenon is a population-level result—it holds in the limit of infinite data. The finite-sample theory is more nuanced. But the population limit identifies the right target and the right test statistic, which the finite-sample theory then makes operational.

13.4 Closing Remarks

The paper of Santoro, Waghmare, and Panaretos is a rare specimen: a result that is simultaneously a contribution to pure mathematics (the theory of Gaussian measures on Hilbert spaces), applied mathematics (functional analysis and operator theory), and statistics (two-sample testing). Its proof is short and technically tight; its implications are broad and practically significant.

The blessing of infinite dimensionality, at least in the form identified here, is not mere metaphor. It is a precise mathematical statement: the RKHS, equipped with the kernel covariance structure and the Gaussian measure, provides a geometric arena in which the hardest problems of nonparametric testing are perfectly resolved. The challenge remains to design finite-sample procedures that approach this ideal, and the follow-up work of Santoro and Panaretos suggests that the gap between the ideal and the attainable is bridgeable.

The interaction between infinite-dimensional geometry, Gaussian measure theory, and statistical inference opened up by this paper is likely to be productive for years to come.

Appendix A

Proof of the Feldman–Hájek Theorem

We sketch the main ideas in the proof of Theorem 4.1.

A.1 Reduction to the Diagonal Case

Without loss of generality, work in the case $\mathbf{m}_1 = \mathbf{m}_2 = 0$ (the mean-shift conditions can be handled separately). Let $\mathbf{C}_1 = \mathbf{C}$ be the reference covariance. Write the eigendecomposition

$$\mathbf{C} = \sum_{j=1}^{\infty} \lambda_j \, e_j \otimes e_j, \tag{A.1}$$

where $\{e_j\}$ is an orthonormal basis and $\lambda_j > 0$ are eigenvalues summing to $\operatorname{trace}(\mathbf{C}) < \infty$.

Any measure $\mathcal{N}(0, \mathbf{C}_2)$ can be written in this basis as $\mathcal{N}(0, \mathbf{D})$ where $\mathbf{D}$ is diagonal (or at least, the relative measures can be studied component by component using the Kakutani product measure theorem).

A.2 Kakutani's Product Theorem

The key tool is Kakutani's theorem on infinite product measures. Consider independent sequences $X_j \sim \mathcal{N}(0, \lambda_j)$ and $Y_j \sim \mathcal{N}(0, \mu_j)$. The product measures $\bigotimes_j \mathcal{N}(0, \lambda_j)$ and $\bigotimes_j \mathcal{N}(0, \mu_j)$ are either equivalent or singular, with the dichotomy governed by:

$$\prod_{j=1}^{\infty} \left(\frac{2\sqrt{\lambda_j \mu_j}}{\lambda_j + \mu_j} \right)^{1/2}, \tag{A.2}$$

the Hellinger affinity of the component Gaussians. The product converges to a positive limit (hence equivalence) if and only if $\sum_j (\sqrt{\mu_j/\lambda_j} - 1)^2 < \infty$; otherwise it converges to zero (hence singularity).

A.3 Connecting to Hilbert–Schmidt

The condition $\sum_j (\sqrt{\mu_j/\lambda_j} - 1)^2 < \infty$ is equivalent (to first order in $\mu_j/\lambda_j - 1$) to $\sum_j (\mu_j/\lambda_j - 1)^2 < \infty$, which is precisely the Hilbert–Schmidt condition $\mathbf{C}_1^{-1/2}\mathbf{C}_2\mathbf{C}_1^{-1/2} - \mathbf{I} \in \mathrm{HS}(\mathcal{H})$. The full Feldman–Hájek theorem follows by combining this spectral analysis with the mean-shift condition.

Appendix B

Operator Norms and the Hilbert–Schmidt Space

B.1 The Hilbert–Schmidt Space

The space $\mathrm{HS}(\mathcal{H})$ of Hilbert–Schmidt operators on $\mathcal{H}$ is itself a Hilbert space under the inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{HS}} = \mathrm{trace}(\mathbf{A}^* \mathbf{B}) = \sum_j \langle \mathbf{A} e_j, \mathbf{B} e_j \rangle, \tag{B.1}$$

for any orthonormal basis $\{e_j\}$. The induced norm is $\|\mathbf{A}\|_{\mathrm{HS}}^2 = \mathrm{trace}(\mathbf{A}^* \mathbf{A})$.

For a self-adjoint operator with eigenvalues $\{\sigma_j(\mathbf{A})\}$:

$$\|\mathbf{A}\|_{\mathrm{HS}}^2 = \sum_j \sigma_j(\mathbf{A})^2, \qquad \|\mathbf{A}\|_{\mathrm{tr}} = \sum_j |\sigma_j(\mathbf{A})|. \tag{B.2}$$

B.2 Trace-Class Operators and Their Role

Trace-class operators play a central role because they are the correct class of covariance operators for Gaussian measures on infinite-dimensional Hilbert spaces. The condition $\mathrm{trace}(\mathbf{C}) < \infty$ ensures that the associated Gaussian process (with covariance operator $\mathbf{C}$) has sample paths in $\mathcal{H}$ almost surely.

The Hilbert–Schmidt condition in the Feldman–Hájek theorem is weaker than the trace-class condition: if $\mathbf{A}$ is trace-class, it is Hilbert–Schmidt, but not conversely. This is why the Feldman–Hájek criterion is non-trivial: it requires the *relative* perturbation to be Hilbert–Schmidt, not the absolute difference.

Appendix C

Background on Locally Compact Polish Spaces

C.1 Polish Spaces

A *Polish space* is a topological space that is separable and completely metrisable. Polish spaces include $\mathbb{R}^d$ and all separable Banach and Hilbert spaces, compact metric spaces, countable discrete spaces, and closed subsets of Polish spaces.

Polish spaces support a well-behaved measure theory: Borel probability measures on Polish spaces are regular (inner and outer regular), and weak convergence of measures is well-defined.

C.2 Local Compactness

A topological space is *locally compact* if every point has a compact neighbourhood. Locally compact Polish spaces include $\mathbb{R}^d$, compact metric spaces, and locally compact groups.

Local compactness is used in the proof of the separation theorem to ensure that the RKHS is dense in $C_0(\mathcal{X})$ (the definition of universality) and that the covariance embedding has the required trace-class properties.

Bibliography

[1] Flyxion, *Kernel Embeddings and the Separation of Measure Phenomenon: A Monograph*, Independent Research monograph, June 2026.

[2] L.V. Santoro, K.G. Waghmare, & V.M. Panaretos, "Kernel embeddings and the separation of measure phenomenon," *Proc. Natl. Acad. Sci. U.S.A.* **123**(23), e2522504123 (2026). `https://doi.org/10.1073/pnas.2522504123`

[3] L.V. Santoro & V.M. Panaretos, "Likelihood ratio tests by kernel Gaussian embedding," *arXiv:2508.07982* (2025).

[4] L.V. Santoro, K.G. Waghmare, & V.M. Panaretos, "Kernel embeddings and the separation of measure phenomenon," *arXiv:2505.04613* (2025).

[5] J. Feldman, "Equivalence and perpendicularity of Gaussian processes," *Pacific J. Math.* **8**, 699–708 (1958).

[6] J. Hájek, "On a property of normal distribution of any stochastic process," *Czechoslovak Math. J.* **8**, 610–618 (1958).

[7] V.I. Bogachev, *Gaussian Measures*, Mathematical Surveys and Monographs, American Mathematical Society (1998).

[8] A.V. Skorokhod, *Integration in Hilbert Space*, Springer (1974).

[9] A. Smola, A. Gretton, L. Song, & B. Schölkopf, "A Hilbert space embedding for distributions," in *Algorithmic Learning Theory*, Springer, 13–31 (2007).

[10] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, & A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.* **13**, 723–773 (2012).

[11] K. Muandet, K. Fukumizu, B. Sriperumbudur, & B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Found. Trends Mach. Learn.* **10**(1–2), 1–141 (2017).

[12] M. Briol, C.J. Oates, M. Girolami, M.A. Osborne, & D. Sejdinovic, "Probabilistic integration: A role in statistical computation?" *Statist. Sci.* **34**(1), 1–22 (2019).

[13] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G.R.G. Lanckriet, & B. Schölkopf, "Hilbert space embedding and characteristic kernels above a Gaussian kernel," *J. Mach. Learn. Res.* **11**, 1517–1561 (2010).

[14] S. Kakutani, "On equivalence of infinite product measures," *Ann. Math.* **49**(1), 214–224 (1948).

[15] G. Da Prato & J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press (1992).

[16] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press (1998).

[17] M.A. Lifshits, *Lectures on Gaussian Processes*, Springer (2012).

[18] B. Schölkopf & A.J. Smola, *Learning with Kernels*, MIT Press (2002).

[19] I. Steinwart & A. Christmann, *Support Vector Machines*, Springer (2008).

[20] J.-F. Briol, F.-X. Briol, & B.K. Sriperumbudur, "Kernel methods for nonparametric hypothesis testing: a review," *Statist. Surveys* (2023).