

The Lift Unification: Why Six Pillars of Probability Fail for the Same Geometric Reason

A Pedagogical Falsification of the Flat Probability Assumption

Bee Rosa Davis

Davis Geometric

bee_davis@alumni.brown.edu

ORCID: 0009-0009-8034-4308

March 2026

Abstract

We demonstrate that six foundational tools of discrete probability — conditional probability, independence, the law of total probability, Bayes’ rule, signal-vs-prior reasoning, and conditional expectation — all fail human intuition through a single unified mechanism. The **Lift**, defined as $\Lambda(A, B) = P(A \cap B) / (P(A) \cdot P(B))$, appears identically in all six frameworks: it governs the non-uniform warping of hypothesis classes under conditioning, the deviation from the product formula under dependence, the hidden multiplier that distorts the tree diagram’s visual grammar, the asymmetric violence of the Bayesian flip, the ratio of signal strength to prior resistance that determines when evidence can overcome a base rate, and the exponential penalty that warps Mossel’s dice paradox from 3 to $3/2$.

Every canonical “paradox” in introductory probability — the mammogram problem, Mossel’s dice paradox, the prosecutor’s fallacy, and the Monty Hall problem — reduces to the same structural cause: the Lift departs significantly from 1 while the pedagogy assumes it equals 1. The product formula $P(A \cap B) = P(A) \cdot P(B)$ is identified as the parallel postulate applied to probability space: exact when the space is flat, and failing with quantifiable, irreducible error when it is not.

This analysis is conducted entirely within standard probability theory in the six falsifications, using no geometric or manifold-theoretic language. The fact that the analysis is forced, by its own internal logic, to conclude with a question about signal, curvature, and capacity suggests that the geometric structure is not an external imposition but the natural completion of a pattern the flat framework can see but cannot name.

Contents

1	Introduction	4
2	Preliminaries	4
3	Falsification 1: Conditional Probability Does Not Rescale Uniformly	5
3.1	The Standard Picture	5
3.2	The Falsification	5
3.3	The Mammogram	6
3.4	The Psychophysical Failure	6
3.5	The Structural Condition	7
4	Falsification 2: Independence Is a Knife-Edge, Not a Baseline	7
4.1	The Standard Picture	7
4.2	The Falsification	7
4.3	The Arithmetic Reason	8
4.4	Connection to Falsification 1	8
5	Falsification 3: The Tree Diagram Hides Multiplicative Violence Behind Additive Bookkeeping	8
5.1	The Standard Picture	8
5.2	The Falsification	9
5.3	Monty Hall	9
5.4	The Mammogram Tree	9
6	Falsification 4: Bayes' Rule Is a Symmetric Coupling Multiplied by an Asymmetric Prior	10
6.1	The Standard Picture	10
6.2	The Falsification	10
6.3	The Mammogram Flip	10
6.4	The Prosecutor's Fallacy	10

6.5	The Cognitive Trap of Standard Grouping	11
6.6	The Violence Predictor	11
7	Falsification 5: The Mammogram Is a Signal Propagating Through a Curved Prior	11
7.1	From Lift to Capacity	11
7.2	The Capacity Interpretation	12
7.3	When $C > 1$: The Test Dominates	12
7.4	The General Principle	13
8	Falsification 6: Mossel’s Dice Paradox and the Curvature of the Conditional Space	13
8.1	The Paradox	13
8.2	Why Everyone Gets It Wrong	14
8.3	The Curvature of the Conditional Space	14
8.4	The Factor of 2	15
9	The Unification	16
10	The Non-Decoupling Theorem: Why This Pattern Exists	17
10.1	Origin	17
10.2	The Theorem	17
10.3	Application to Probability	17
10.4	The Governing Equation	18
10.5	Scope	19
11	Conclusion	19
12	Acknowledgments	20

1 Introduction

MIT’s 6.042 (Mathematics for Computer Science) is one of the most widely viewed mathematics courses in the world, with millions of views on YouTube through MIT OpenCourseWare. In a representative lecture on probability, Professor Ankur Moitra introduces conditional probability, independence, the law of total probability, and Bayes’ rule — four tools that form the backbone of probabilistic reasoning across science, engineering, medicine, and law [1].

The lecture is excellent. The algebra is exact. The examples are well-chosen. And yet the lecture contains a systematic pedagogical gap: every tool is presented with a visualization or default intuition that fails precisely when the tool matters most.

The dartboard picture for conditional probability implies uniform rescaling. It isn’t uniform. The product formula for independence is presented as the simple baseline. It is the rarest special case. The tree diagram for total probability draws branches proportional to prior weights. The posteriors are not proportional to prior weights. Bayes’ rule is presented as a symmetric algebraic flip. The flip can be arbitrarily violent.

This paper demonstrates that all six failures share a single cause. The quantity

$$\Lambda(A, B) = \frac{P(A \cap B)}{P(A) \cdot P(B)} \tag{1}$$

known in data mining as **Lift** and in probability as the ratio form of the Bayes factor, appears identically in all four frameworks. It controls the magnitude of every departure from naive intuition. When $\Lambda \approx 1$, all four tools behave as the standard pedagogy suggests. When Λ departs significantly from 1, all four tools produce results that humans — including trained physicians and at least one unnamed MIT faculty member — systematically get wrong.

The paper proceeds as six falsifications, each targeting one pillar of the lecture, followed by a unification that reveals the common structure and a discussion of the mathematical framework that explains why the pattern exists.

2 Preliminaries

We work entirely within discrete probability on finite sample spaces, using standard notation. Let S be a finite sample space with probability measure P . Events are subsets of S . All probabilities are non-negative and sum to 1 over S .

Definition 1 (Lift). *For events A, B with $P(A) > 0$ and $P(B) > 0$, the **Lift** is*

$$\Lambda(A, B) = \frac{P(A \cap B)}{P(A) \cdot P(B)} \tag{2}$$

Remark 2. *The Lift has several equivalent forms:*

$$\Lambda(A, B) = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)} \quad (3)$$

The first equality follows from the definition of conditional probability. The second follows from symmetry of the numerator $P(A \cap B) = P(B \cap A)$.

Remark 3. *The Lift is related to the covariance of indicator random variables by*

$$\Lambda(A, B) = 1 + \frac{\text{Cov}(\mathbb{I}_A, \mathbb{I}_B)}{P(A) \cdot P(B)} \quad (4)$$

where $\text{Cov}(\mathbb{I}_A, \mathbb{I}_B) = P(A \cap B) - P(A) \cdot P(B)$.

Remark 4. *The logarithm of the Lift is the Pointwise Mutual Information (PMI) from information theory: $\text{PMI}(A, B) = \log \Lambda(A, B)$.*

3 Falsification 1: Conditional Probability Does Not Rescale Uniformly

3.1 The Standard Picture

The standard visualization of $P(A|B) = P(A \cap B)/P(B)$ is the dartboard: the sample space S is a rectangle, events A and B are overlapping regions, and conditioning on B is described as “zooming in” on region B and renormalizing so that probabilities within B sum to 1 [1].

The visual implication is that this zoom preserves proportions — the relative sizes of sub-regions within B remain the same after conditioning, just scaled up by the constant factor $1/P(B)$.

3.2 The Falsification

At the level of atomic outcomes, the rescaling is indeed uniform: every $\omega \in B$ is rescaled by $1/P(B)$. However, human reasoning does not operate at the atomic level. Physicians, engineers, and students reason about *hypothesis classes* — composite events like “sick” or “healthy” that partition the conditioned space.

At the level of hypothesis classes, the rescaling is governed by the Lift. For any sub-event $H \subseteq S$ with $P(H) > 0$:

$$\frac{P(H|B)}{P(H)} = \frac{P(B|H)}{P(B)} = \Lambda(H, B) \quad (5)$$

This factor varies across hypothesis classes whenever $P(B|H)$ is not constant — which is precisely when conditioning is informative.

3.3 The Mammogram

Consider the standard example from Kahneman [2]. Population: 1000 women. Base rate of breast cancer: 1% (10 sick, 990 healthy). Mammogram sensitivity: 90%. False positive rate: 10%.

Before conditioning on a positive test:

- Sick : Healthy = 10 : 990 = 1 : 99

After conditioning on a positive test (108 total positives: 9 true, 99 false):

- Sick : Healthy = 9 : 99 = 1 : 11

The prior odds were warped by the likelihood ratio:

$$\text{Posterior odds} = \text{Prior odds} \times \frac{P(B|\text{Sick})}{P(B|\text{Healthy})} = \frac{1}{99} \times \frac{0.9}{0.1} = \frac{9}{99} = \frac{1}{11} \quad (6)$$

The Lift for the Sick hypothesis:

$$\Lambda(\text{Sick}, \text{Positive}) = \frac{P(\text{Pos}|\text{Sick})}{P(\text{Pos})} = \frac{0.9}{0.108} = 8.33 \quad (7)$$

The Lift for the Healthy hypothesis:

$$\Lambda(\text{Healthy}, \text{Positive}) = \frac{P(\text{Pos}|\text{Healthy})}{P(\text{Pos})} = \frac{0.1}{0.108} = 0.926 \quad (8)$$

The Sick hypothesis is amplified by 8.33 $\times$ while the Healthy hypothesis is compressed to 0.926 $\times$. This is a 9 $\times$ differential — the likelihood ratio. The dartboard picture shows none of this.

3.4 The Psychophysical Failure

A fair counterargument: if the dartboard were drawn strictly to scale, the area of (Sick $\cap$ Positive) relative to the total area of (Positive) would be exactly 8.3%. The geometry is technically accurate.

But the human visual cortex cannot compute area-to-area ratios of highly asymmetric, overlapping slivers — especially when the scale difference is 100:1. The visualization asks the brain to perform a Euclidean estimation task it is biologically incapable of executing, which forces the brain to substitute a heuristic: anchoring on the 90% sensitivity and ignoring the denominator entirely. This is Kahneman's *base-rate neglect* [2].

Gigerenzer's natural frequency research confirms this diagnosis [3]. When physicians are forced to use explicit population counts (1000 people, 10 sick, 9 true positives, 99 false positives), success rates jump from roughly 15% to over 70%. This works precisely because

integer counting makes the physical magnitude of the covariance visible within the computation itself. When you see 9 true positives physically drowning in 99 false positives, you are visually processing the Lift. The dartboard hides it. Gigerenzer’s fix succeeds because it corrects exactly the flaw identified here.

3.5 The Structural Condition

The distortion becomes catastrophically counterintuitive when the false positive rate exceeds the base rate: $P(B|H_2) > P(H_1)$. In the mammogram, $\text{FPR} = 0.10$ and $\text{base rate} = 0.01$. The ratio is 10. This single number — the signal-to-noise ratio of the test relative to the prior — predicts in advance how badly the “uniform zoom” intuition will fail.

4 Falsification 2: Independence Is a Knife-Edge, Not a Baseline

4.1 The Standard Picture

Independence is defined as $P(A \cap B) = P(A) \cdot P(B)$ and presented as the “simple case” — two dice in separate rooms, two experiments that don’t talk to each other [1]. Dependence is then the complication you check for.

4.2 The Falsification

Using Equation 4, independence is equivalent to $\Lambda(A, B) = 1$, which is equivalent to $\text{Cov}(\mathbb{I}_A, \mathbb{I}_B) = 0$.

We compute the covariance on three pairs of events from Moitra’s own lecture. Let the sample space be $\{1, \dots, 6\}^2$ (two dice, 36 equally likely outcomes). Define:

- A : x_1 is odd
- B : $x_2 \leq 3$
- C : $x_1 - x_2 = 3$
- D : $x_1 - x_2 = 4$

$$\text{Cov}(A, B) = \frac{9}{36} - \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} - \frac{1}{4} = 0 \quad \Lambda = 1 \quad (\text{independent}) \quad (9)$$

$$\text{Cov}(A, C) = \frac{1}{36} - \frac{1}{2} \cdot \frac{3}{36} = -\frac{1}{72} \quad \Lambda = 0.667 \quad (\text{negatively coupled}) \quad (10)$$

$$\text{Cov}(A, D) = \frac{1}{36} - \frac{1}{2} \cdot \frac{2}{36} = 0 \quad \Lambda = 1 \quad (\text{independent}) \quad (11)$$

Events C and D have identical structure — both constrain $x_1 - x_2$, both share x_1 with event A . Yet one couples to A and the other doesn't.

4.3 The Arithmetic Reason

When $x_1 - x_2 = 3$, the qualifying pairs are $(4, 1), (5, 2), (6, 3)$. Only one has odd x_1 : the pair $(5, 2)$. That is $1/3$ — less than the base rate of $1/2$ for odd. The constraint is asymmetric across odd and even values of x_1 .

When $x_1 - x_2 = 4$, the qualifying pairs are $(5, 1), (6, 2)$. One has odd x_1 : the pair $(5, 1)$. That is $1/2$ — matching the base rate exactly. The coupling vanishes by arithmetic coincidence.

Changing the constraint by a single integer — from 3 to 4 — flips the covariance from nonzero to exactly zero. Independence is a knife-edge condition, not a robust property. Zero coupling is the anomaly. Coupling is the default.

4.4 Connection to Falsification 1

The Lift-covariance identity (Equation 4) connects this directly to the mammogram:

$$\Lambda(\text{Sick, Positive}) = 1 + \frac{0.00792}{0.01 \times 0.108} = 1 + 7.33 = 8.33 \quad (12)$$

The normalized covariance of 7.33 is the Lift minus 1 — exactly the distortion factor from Falsification 1. The same quantity governs both the failure of the dartboard and the failure of the product formula.

5 Falsification 3: The Tree Diagram Hides Multiplicative Violence Behind Additive Bookkeeping

5.1 The Standard Picture

The law of total probability states $P(A) = \sum_i P(A|B_i)P(B_i)$ where the B_i are disjoint and cover S . Moitra visualizes this as a tree: you descend a branch with probability $P(B_i)$, then event A occurs with conditional probability $P(A|B_i)$, and you sum across branches [1].

The visual grammar of the tree draws branches proportional to $P(B_i)$. The implication: thick branches dominate the total, thin branches are negligible.

5.2 The Falsification

The posterior contribution of each branch is not proportional to its prior weight. By Bayes’ rule:

$$P(B_i|A) = \Lambda(B_i, A) \times P(B_i) \quad (13)$$

The posterior is the prior weight *warped by the Lift*. The tree shows you $P(B_i)$. It does not show you the multiplier $\Lambda(B_i, A)$.

5.3 Monty Hall

In Moitra’s presentation: B_1 = “picked the prize door” with $P(B_1) = 1/3$, B_2 = “didn’t” with $P(B_2) = 2/3$. Under the “always switch” strategy, $P(\text{win}|B_1) = 0$ and $P(\text{win}|B_2) = 1$.

$$\Lambda(B_1, \text{win}) = \frac{P(\text{win}|B_1)}{P(\text{win})} = \frac{0}{2/3} = 0 \quad (14)$$

$$\Lambda(B_2, \text{win}) = \frac{P(\text{win}|B_2)}{P(\text{win})} = \frac{1}{2/3} = \frac{3}{2} \quad (15)$$

The tree looks like two branches — one carrying $1/3$ of the weight and one carrying $2/3$. But branch B_1 contributes *nothing* to the posterior. Branch B_2 contributes *everything*. The democratic picture is a lie.

5.4 The Mammogram Tree

Partition into B_1 = sick ($P = 0.01$) and B_2 = healthy ($P = 0.99$). The tree draws B_1 as a tiny thread and B_2 as a massive trunk.

$$\Lambda(B_1, \text{Positive}) = 8.33 \quad (16)$$

$$\Lambda(B_2, \text{Positive}) = 0.926 \quad (17)$$

The tree draws B_1 as tiny, but the Lift multiplies its influence by $8.33\times$. The tree draws B_2 as massive, but the Lift compresses it. The additive visual structure of the tree entirely hides the multiplicative violence of the Lift.

This is why physicians’ intuition fails: their mental model relies on the visual thickness of the prior branch (1%), totally blind to the fact that the Lift is acting as an $8\times$ multiplier on that specific branch. A branch that looks negligible in the tree can dominate the posterior. A branch that looks dominant can become nearly irrelevant.

6 Falsification 4: Bayes' Rule Is a Symmetric Coupling Multiplied by an Asymmetric Prior

6.1 The Standard Picture

Bayes' rule is presented as $P(B|A) = P(A|B) \cdot P(B)/P(A)$ — a symmetric algebraic flip of the conditional. The proof is one line. Moitra mentions the historical footnote about trying to prove God [1].

6.2 The Falsification

The Lift decomposition reveals the hidden structure of the flip. Using Equation 3:

$$P(A|B) = \Lambda(A, B) \times P(A) \tag{18}$$

$$P(B|A) = \Lambda(A, B) \times P(B) \tag{19}$$

Same Lift. Different priors. That is the entire content of Bayes' rule.

When $P(A) \approx P(B)$, the flip is gentle. When $P(A)$ and $P(B)$ are wildly different, the flip is violent. The algebra is identical in both cases. The psychological experience is not.

6.3 The Mammogram Flip

A = test positive, B = actually sick. $\Lambda(A, B) = 8.33$.

$$P(A|B) = 8.33 \times 0.108 = 0.9 \tag{20}$$

$$P(B|A) = 8.33 \times 0.01 = 0.083 \tag{21}$$

Same Lift. One direction gives 90%. The other gives 8.3%. The asymmetry is not in the coupling — it is entirely in the priors.

6.4 The Prosecutor's Fallacy

$P(\text{evidence}|\text{guilty})$ is not $P(\text{guilty}|\text{evidence})$. A forensic test matching 1 in 10,000 means $P(\text{match}|\text{innocent}) = 0.0001$. A prosecutor presents this as “0.01% chance of innocence.” But if the suspect pool is a city of 1,000,000, there are 100 innocent matches. $P(\text{innocent}|\text{match})$ could be as high as 99%.

Same Lift. Different priors. Opposite conclusions.

6.5 The Cognitive Trap of Standard Grouping

The standard formula groups the numerator first: $P(B|A) = [P(A|B) \cdot P(B)]/P(A)$. Students compute the joint $P(A \cap B)$, then divide by the marginal. This standard sequence physically breaks the Lift apart — it blends the symmetric coupling strength with the asymmetric prior before the student ever sees them as separate quantities.

Shifting the parentheses — $P(B|A) = [P(A|B)/P(A)] \times P(B) = \Lambda \times P(B)$ — makes the cognitive trap visible. The student now sees two independent components: a symmetric coupling term (Lift) and an asymmetric prior. The standard formula hides this decomposition.

6.6 The Violence Predictor

The ratio of the two posteriors is:

$$\frac{P(A|B)}{P(B|A)} = \frac{P(A)}{P(B)} \quad (22)$$

The prior ratio alone determines how different the two directions of Bayes' rule will feel. In the mammogram, $P(A)/P(B) = 0.108/0.01 = 10.8$. This number tells you the flip will produce answers differing by roughly an order of magnitude — computable before touching Bayes' rule.

7 Falsification 5: The Mammogram Is a Signal Propagating Through a Curved Prior

7.1 From Lift to Capacity

The previous four falsifications established that the Lift Λ governs every departure from intuition. But the Lift itself is a *symptom*, not a cause. We now ask: what *generates* extreme Lift values?

Return to the mammogram. The Lift for the Sick hypothesis is:

$$\Lambda(\text{Sick}, \text{Positive}) = \frac{P(\text{Pos}|\text{Sick})}{P(\text{Pos})} = \frac{0.9}{0.108} = 8.33 \quad (23)$$

This number has two components: a numerator (the signal strength of the test, 0.9) and a denominator (the marginal probability of a positive test, 0.108). The denominator is where the base rate hides:

$$P(\text{Pos}) = P(\text{Pos}|\text{Sick}) \cdot P(\text{Sick}) + P(\text{Pos}|\text{Healthy}) \cdot P(\text{Healthy}) = 0.9 \times 0.01 + 0.1 \times 0.99 = 0.108 \quad (24)$$

When the base rate $P(\text{Sick})$ is small, the first term is negligible and $P(\text{Pos}) \approx \text{FPR} \times P(\text{Healthy}) \approx \text{FPR}$. The denominator of the Lift is dominated by the false positive rate acting on the vast healthy population.

Define two quantities:

- **Signal strength** τ : the test’s discriminative power. A natural measure is the log-likelihood ratio: $\tau = \ln \left(\frac{P(\text{Pos}|\text{Sick})}{P(\text{Pos}|\text{Healthy})} \right) = \ln \left(\frac{0.9}{0.1} \right) = \ln(9) \approx 2.20$
- **Prior resistance** K : a measure of how asymmetric the base rates are. A natural measure is $K = \ln \left(\frac{P(\text{Healthy})}{P(\text{Sick})} \right) = \ln \left(\frac{0.99}{0.01} \right) = \ln(99) \approx 4.60$

(We use natural logarithm throughout; the ratio τ/K is base-independent since the base cancels.)

In log-odds space, the posterior log-odds equal the prior log-odds plus the log-likelihood ratio:

$$\underbrace{\ln \frac{P(\text{Sick}|\text{Pos})}{P(\text{Healthy}|\text{Pos})}}_{\text{posterior log-odds}} = \underbrace{\ln \frac{P(\text{Sick})}{P(\text{Healthy})}}_{\text{prior log-odds}} + \underbrace{\ln \frac{P(\text{Pos}|\text{Sick})}{P(\text{Pos}|\text{Healthy})}}_{\tau} \quad (25)$$

The prior log-odds are $\ln(0.01/0.99) = -4.60 = -K$. The signal adds $\tau = 2.20$. The posterior log-odds are $-4.60 + 2.20 = -2.40$, corresponding to posterior probability ≈ 0.083 .

7.2 The Capacity Interpretation

The test “moves” the log-odds by τ . But the prior “resists” by K . The net movement is $\tau - K$. When $\tau < K$, the test cannot overcome the prior, and the posterior remains on the same side as the prior (below 50%). When $\tau > K$, the test flips the posterior past 50%.

Define the **capacity** of the test relative to the prior:

$$C = \frac{\tau}{K} \quad (26)$$

In the mammogram: $C = 2.20/4.60 = 0.478$. The test has less than half the signal strength needed to overcome the prior curvature. Its capacity is less than 1. It cannot flip the posterior. It can only nudge it — from 1% to 8.3%.

This is not a property of the test alone (90% sensitivity sounds impressive) or the prior alone (1% base rate sounds small but manageable). It is a property of their *ratio*. The test is strong. The prior is stronger. $C < 1$ tells you this instantly, before any computation.

7.3 When $C > 1$: The Test Dominates

Suppose the base rate were 10% instead of 1%. Then $K = \ln(0.9/0.1) = 2.20$ and the signal-to-resistance ratio is $\tau/K = 2.20/2.20 = 1.0$. The test has exactly enough signal to

match the prior resistance. The posterior would be 50% — maximum uncertainty.

If the base rate were 20%, then $K = \ln(0.8/0.2) = 1.39$ and $\tau/K = 2.20/1.39 = 1.58$. Now the ratio exceeds 1. The test overwhelms the prior. The posterior probability of being sick given a positive test would be roughly 69%. In this regime, the physician’s intuition (“positive test means probably sick”) would be approximately correct.

The Lift captures the magnitude of the distortion. $C = \tau/K$ captures *why*. It decomposes the Lift into signal (what the evidence contributes) and curvature (what the prior resists with). The flat probability framework gives you $\Lambda = 8.33$ and says “that’s the number.” The capacity framework says: the number is 8.33 *because* the signal-to-curvature ratio is 0.478, and any test with $C < 1$ will produce a Lift that amplifies the rare hypothesis without being strong enough to make it probable.

7.4 The General Principle

The Lift Λ tells you *how much* the intuition fails. The capacity $C = \tau/K$ tells you *why* it fails — the signal couldn’t overcome the curvature. Any conditional probability problem where $C \ll 1$ will produce extreme Lifts, counterintuitive posteriors, and systematic errors in human reasoning.

The quantity C is computable from the structure of the problem before any Bayesian computation. It is the answer to the question posed at the end of Falsification 4: a single number that tells you in advance whether your intuition is safe.

8 Falsification 6: Mossel’s Dice Paradox and the Curvature of the Conditional Space

8.1 The Paradox

Mossel’s dice paradox, presented at the end of the lecture [6], is the most striking example of conditional probability defying expert intuition.

Experiment: Roll a fair die repeatedly until you get a 6. **Question:** Conditioned on all rolls being even, what is the expected number of rolls?

The “obvious” answer is 3. The reasoning: if all rolls are even, then only 2, 4, and 6 can appear. A 6 is 1/3 of the even faces. By the mean time to failure formula, the expected number of rolls is $1/(1/3) = 3$.

A math professor argued for months that the answer was 3. Mossel himself initially thought he had made an algebraic error on the board.

The actual answer is 3/2.

8.2 Why Everyone Gets It Wrong

The “obvious” argument assumes that conditioning on “all rolls are even” is equivalent to replacing the die with a flat three-sided die showing $\{2, 4, 6\}$, each with probability $1/3$. This is the dartboard assumption from Falsification 1: conditioning just rescales the space uniformly.

But the conditional space is not a flat three-sided die. Here is why.

Consider the unconditioned experiment. The probability of the sequence $(e_1, e_2, \dots, e_k)$ where $e_k = 6$ and all previous $e_i \in \{2, 4\}$ is:

$$P(\text{sequence of length } k) = \left(\frac{2}{6}\right)^{k-1} \times \frac{1}{6} = \frac{2^{k-1}}{6^k} \quad (27)$$

The probability that all rolls are even AND the sequence has length k is this same quantity. The probability that all rolls are even is the sum over all k :

$$P(\text{all even}) = \sum_{k=1}^{\infty} \frac{2^{k-1}}{6^k} = \frac{1/6}{1 - 2/6} = \frac{1/6}{4/6} = \frac{1}{4} \quad (28)$$

Now the conditional probability of length k given all even:

$$P(\text{length } k \mid \text{all even}) = \frac{2^{k-1}/6^k}{1/4} = \frac{4 \cdot 2^{k-1}}{6^k} = \frac{2^{k+1}}{6^k} = 2 \cdot \left(\frac{1}{3}\right)^k \quad (29)$$

Check: for $k = 1$, $P = 2/3$. For $k = 2$, $P = 2/9$. For $k = 3$, $P = 2/27$. These sum to 1 (geometric series with ratio $1/3$ starting at $2/3$).

The expected length:

$$E[k \mid \text{all even}] = \sum_{k=1}^{\infty} k \cdot 2 \cdot \left(\frac{1}{3}\right)^k = 2 \cdot \frac{1/3}{(1 - 1/3)^2} = 2 \cdot \frac{1/3}{4/9} = 2 \cdot \frac{3}{4} = \frac{3}{2} \quad (30)$$

8.3 The Curvature of the Conditional Space

The “flat three-sided die” argument assumes that in the conditional space, each roll is an independent $1/3$ chance of being a 6, giving geometric distribution with mean 3. But the actual conditional distribution over sequence lengths is $P(k) = 2 \cdot (1/3)^k$, which is *not* the geometric distribution with parameter $1/3$.

In a geometric distribution with parameter $p = 1/3$, we would have $P(k) = (2/3)^{k-1} \cdot (1/3)$. For $k = 1$: $P = 1/3$. For $k = 2$: $P = 2/9$.

In the actual conditional distribution: $P(1) = 2/3$. $P(2) = 2/9$.

The conditional distribution assigns *twice* the probability to $k = 1$ compared to the flat model. It is massively biased toward short sequences.

Why? Because the event “all rolls are even” becomes exponentially less likely as the sequence gets longer. A length-1 sequence has probability $1/2$ of being all even (just need one even roll). A length-2 sequence has probability $1/4$. Length 3 has probability $1/8$. Conditioning on “all even” doesn’t just restrict which faces appear — it exponentially penalizes long sequences.

This is non-uniform rescaling, exactly as in Falsification 1. The Lift for short sequences is much larger than the Lift for long sequences:

$$\Lambda(\text{length } k, \text{ all even}) = \frac{P(\text{length } k \mid \text{all even})}{P(\text{length } k)} = \frac{2 \cdot (1/3)^k}{(5/6)^{k-1} \cdot (1/6)} \quad (31)$$

For $k = 1$: $\Lambda = (2/3)/(1/6) = 4$. For $k = 2$: $\Lambda = (2/9)/(5/36) = 8/5 = 1.6$. For $k = 3$: $\Lambda = (2/27)/(25/216) = 16/25 = 0.64$.

Short sequences are amplified. Long sequences are compressed. The Lift decreases with k . The conditional space is warped toward brevity.

8.4 The Factor of 2

Everyone expects 3. The answer is $3/2$. The ratio between the “obvious” answer and the true answer is exactly 2.

This factor invites comparison with the Double Cover Principle [10]. The “obvious” model treats the conditional space as a flat die: three faces, equal probability, geometric distribution with mean 3. The actual conditional space is warped by the exponential penalty on sequence length, producing mean $3/2$.

In the Double Cover framework, the identity $S + d^2 = 1$ separates what is preserved under a symmetry operation from what is lost. The flat model assumes full preservation ($S = 1$): conditioning restricts which faces appear but preserves the distributional symmetry among them. The actual conditioning operation breaks this symmetry ($d^2 > 0$): longer sequences are exponentially penalized, so the distribution is no longer symmetric among the surviving faces.

Whether the factor of 2 is precisely the covering degree of $S^1 \rightarrow \mathbb{RP}^1$ — as it is in plasma confinement (Troyon limit) and quantum state overlaps (spin-1/2 sameness) — or merely a suggestive coincidence remains an open question. What is established is the mechanism: the conditional space has internal structure that the flat model cannot represent, and this structure compresses the expectation by a factor that the “uniform rescaling” assumption from Falsification 1 would never predict.

9 The Unification

Across all six falsifications, the same structural pattern repeats: a ratio departs from 1, the standard pedagogy assumes it equals 1, and the result is a specific, quantifiable failure of intuition.

Framework	Role of Lift / Capacity	Failure mode
Conditional probability	Warp factor on hypotheses	Dartboard implies uniform zoom
Independence	$\Lambda = 1$ is the definition	Product formula fails
Total probability	Hidden multiplier on branches	Tree implies proportional contribution
Bayes' rule	Symmetric coupling, asymmetric flip	Confusion of the inverse
Signal vs. prior	Signal-to-resistance ratio < 1	Test cannot overcome prior resistance
Conditional expectation	Conditioning warps toward short paths	Flat die model ignores exponential penalty

The six canonical “paradoxes” of introductory probability — the mammogram, the dice coupling, the Monty Hall problem, the prosecutor’s fallacy, base-rate neglect, and Mossel’s dice paradox — are not six different problems. They are one problem wearing six different outfits.

Proposition 5 (Three-Parameter Characterization). *The departure of probabilistic reasoning from naive intuition is completely characterized by three quantities:*

1. The **Lift** $\Lambda(A, B) = P(A \cap B) / (P(A) \cdot P(B))$, which controls the magnitude of distortion under conditioning, the deviation from the product formula, and the hidden multiplier in the tree.
2. The **prior ratio** $P(A) / P(B)$, which controls the asymmetric violence of the Bayesian flip.
3. The **signal-to-resistance ratio** τ / K , which measures whether the signal strength of the evidence can overcome the resistance of the prior. When $\tau / K < 1$, evidence is insufficient to overturn the prior, and intuition systematically overestimates posterior probability.

When all three are near 1, all six tools behave as the standard pedagogy suggests. When any departs significantly from 1, intuition fails in a specific, predictable, quantifiable way.

10 The Non-Decoupling Theorem: Why This Pattern Exists

The previous sections demonstrated *that* the flat probability framework fails in predictable ways. This section explains *why*.

10.1 Origin

The connection between the product formula and the parallel postulate was discovered through a geometric reinterpretation of the pigeonhole principle, following the MIT 6.042 curriculum [1]. The classical pigeonhole principle — more items than containers implies sharing — was recast as a capacity argument on a Riemannian manifold: compactness of the rotation group implies finite distinguishable states, and sufficient path length implies collision. The critical observation was that the geometric pigeonhole principle requires path-dependence, which requires nonzero curvature, which is precisely the condition under which the product formula (the “parallel postulate of probability”) fails.

10.2 The Theorem

This observation was formalized as the Davis Non-Decoupling Theorem [9]:

Theorem 6 (Non-Decoupling, informal statement). *A system admitting a faithful description as a principal fiber bundle with connection ω can be modeled by decoupled (connection-free) components if and only if the curvature 2-form $F = d\omega + \omega \wedge \omega$ vanishes identically. When curvature is nonzero:*

1. *Any flat (decoupled) model incurs irreducible structural error bounded below by the Yang–Mills action $\int \|F\|^2$.*
2. *The error concentrates where the curvature $\|F\|$ is maximal.*
3. *When the bundle carries non-trivial characteristic classes, no flat connection exists — decoupled models are not merely inaccurate but categorically impossible.*

The accumulated error is called the **holonomy debt**: the residual rotation that parallel transport accumulates around a closed loop on a curved space.

10.3 Application to Probability

Every falsification in this paper is an instance of the Non-Decoupling Theorem applied to probability:

Falsification 1: Conditioning on B is parallel transport from the full space S to the subspace B . On a flat space, transport preserves proportions. On a curved space, it warps them by the Lift.

Falsification 2: The product formula is the flat connection. Independence is zero curvature. The covariance is the curvature tensor — it measures the holonomy debt of assuming flatness.

Falsification 3: The tree diagram is an atlas of coordinate charts on the probability space. On a flat space, the charts combine additively. On a curved space, each chart picks up a multiplicative correction (the Lift) from the curvature.

Falsification 4: Bayes’ rule is parallel transport in one direction followed by transport in the reverse direction. On a flat space, the round trip is trivial. On a curved space, the holonomy debt is the prior ratio $P(A)/P(B)$.

Falsification 5: The signal-to-resistance ratio decomposes the Lift into signal and resistance. It measures whether the evidence can overcome the geometric resistance of the prior. When the ratio is less than 1, the resistance wins.

Falsification 6: Mossel’s paradox arises because conditioning on “all even” warps the space non-uniformly toward short sequences. The flat model assumes the conditional space inherits the geometry of the original space. The Double Cover Principle [10] identifies the factor-of-2 error as the covering degree between the holonomy space (S^1) and the sameness space ($\mathbb{RP}^1$).

10.4 The Governing Equation

The Davis Field Equations [8] provide the governing relation for all of these phenomena:

$$C = \frac{\tau}{K} \tag{32}$$

where C is the capacity to propagate signal across a structured space, τ is the tolerance budget (signal strength, discriminative power, evidence), and K is the geometric impedance — the holonomy debt of the specific bundle being traversed.

A critical clarification: K is not a single algebraic expression. It is a geometric quantity that *projects* differently onto each operation:

- In Bayesian updating (Falsification 5), K projects as the log-prior ratio $\ln(P(H_2)/P(H_1))$, measuring the resistance of asymmetric base rates to evidence.
- In joint distributions (Falsification 2), K projects as the pointwise mutual information $-\ln \Lambda(A, B)$, measuring the resistance of the dependence structure to the product formula.
- In conditional expectation (Falsification 6), K projects as the exponential penalty that conditioning imposes on sequence length.

A 50/50 base rate ($\ln(P(H_2)/P(H_1)) = 0$) does not guarantee independence — it only means the *prior-asymmetry projection* of K vanishes. The joint distribution can still carry curvature through its dependence structure. All of classical probability is the limit in which *every* projection of K vanishes simultaneously — the space is flat in every direction, not merely in one.

When $K = 0$ across all projections, capacity is infinite. The product formula works. Independence holds. The dartboard rescales uniformly. Bayes’ rule flips symmetrically. Conditioning preserves structure.

When $K > 0$ in any projection, capacity is finite along that direction. The corresponding tool acquires a correction term proportional to the curvature. The Lift departs from 1. The product formula fails. The tree warps. The flip becomes violent.

10.5 Scope

The Non-Decoupling Theorem and the governing equation $C = \tau/K$ have been validated across eleven branches of the Davis Field Equations program, including plasma confinement [12], pharmacokinetic drug delivery [11], viral surveillance [8], and language model architecture. The present paper adds a twelfth branch: the geometry of probability itself.

The flat probability framework taught in every introductory course is the $K = 0$ special case. The six falsifications demonstrate that the corrections arising from $K > 0$ are not exotic mathematical curiosities — they are the structural cause of the most famous “paradoxes” in the field.

11 Conclusion

We have demonstrated that six foundational tools of discrete probability fail human intuition through a single unified mechanism: the Lift $\Lambda(A, B)$ departs from 1 because the underlying probability space has nonzero curvature that the flat pedagogical framework cannot represent.

The signal-to-resistance ratio τ/K provides the structural explanation: every failure occurs when the signal strength of the evidence is insufficient to overcome the geometric impedance of the space. The impedance projects differently onto each operation — as prior asymmetry in Bayesian updating, as mutual information in joint distributions, as exponential penalty in conditional expectation — but the governing relation is the same.

The Non-Decoupling Theorem provides the mathematical foundation: on any space with nonzero curvature, the product formula (independence assumption) incurs irreducible error bounded below by the curvature. The flat probability framework taught in every introductory course is the $K = 0$ special case of a richer geometric theory.

This paper was written entirely within the vocabulary of standard probability theory, deliberately avoiding geometric or manifold-theoretic language in the six falsifications. The fact that the analysis was forced, by its own internal logic, to conclude with a governing equation involving signal, curvature, and capacity — without ever invoking those words until the final sections — suggests that the geometric structure is not an external imposition. It is the natural completion of the pattern that the flat framework can see but cannot name.

The ratios were always trying to tell us something. Now they have a name. Now they have a theorem. Now they have an equation.

12 Acknowledgments

The lecture analyzed is from MIT 6.042 (Mathematics for Computer Science), taught by Professor Ankur Moitra and freely available through MIT OpenCourseWare. The author is grateful for the open availability of this exceptional course material, and notes that the lecture is excellent — the falsifications target the *framework*, not the instructor. Thank you Professor Moitra for teaching the next generation of bright young minds.

References

- [1] Moitra, A. (2024). MIT 6.042: Mathematics for Computer Science. Lecture on Probability, Conditioning, and Independence. MIT OpenCourseWare. <https://youtu.be/Z1zgrgstVAw>
- [2] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- [3] Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704.
- [4] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- [5] Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd ed.). Springer.
- [6] Mossel, E., & Tamuz, O. (2010). Dice paradox — conditional expectation. *Personal communication / lecture demonstration*.
- [7] Berry, M. V. (1984). Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London A*, 392(1802), 45–57.
- [8] Davis, B. R. (2025). The Field Equations of Semantic Coherence: A Geometric Theory of Meaning, Curvature, and Reasoning in Transformer Architectures. *Zenodo*. <https://doi.org/10.5281/zenodo.17771796>
- [9] Davis, B. R. (2026). The Davis Non-Decoupling Theorem: Why Every Flat Model of a Curved System Fails, and the Topological Size of the Debt. *Zenodo*. <https://doi.org/10.5281/zenodo.18754646>
- [10] Davis, B. R. (2026). The Double Cover Principle: Every Circle Carries Two Circumferences. *Zenodo*. <https://doi.org/10.5281/zenodo.18895462>

- [11] Davis, B. R. (2026). The Geometry of Delivery: A Uniqueness Theorem for Section Coherence over Stratified Barrier Bundles. *Zenodo*. <https://doi.org/10.5281/zenodo.19321978>
- [12] Davis, B. R. (2026). The Spectral Geometry of Plasma Confinement: A Davis Field Equations Framework for Fusion Stability, Transport Bottlenecks, and Cross-Domain Universality. *Zenodo*. <https://doi.org/10.5281/zenodo.18969038>

Author: Bee Rosa Davis — bee_davis@alumni.brown.edu | ORCID: 0009-0009-8034-4308

License: Creative Commons Attribution 4.0 International (CC BY 4.0)

Keywords: conditional probability, independence, Bayes' rule, law of total probability, Lift, covariance, base-rate neglect, prosecutor's fallacy, mammogram problem, Monty Hall, Mossel's dice paradox, Double Cover Principle, Non-Decoupling Theorem, signal-to-curvature ratio, capacity, Davis Field Equations, pigeonhole principle, parallel postulate, holonomy debt, pedagogical falsification