Learning from Data&Prior

Carlos C. Rodríguez

405 Main Street,
New York, NY10044.

Abstract

During the last twenty years the Bayesian approach to statistical inference has been increasing in popularity due to its success in solving practical problems, often by the Monte Carlo method–although not everyone is (or should be) persuaded by the use of subjective priors on arbitrary parameterizations. The problem with Bayesianism are the priors. There is still no general theory for choosing the prior. The purpose of this paper is to provide such a theory. This is made possible by a new geometrization of the concept of ignorance. At the same time this forces a re-evaluation of the meaning of data, of prior, and indeed the meaning of meaning itself. It turns out to be surprisingly simple: Probability is meaning.

1 Introduction: Probability is Objective Meaning

The statement "There is no data in the vacuum" is not a metaphysical puzzle; it is a strict geometric truth. A label x (a number, a pixel, a coordinate) in isolation is devoid of meaning. It only acquires meaning when it is mapped to a probability distribution p that generated it. Therefore, Probability is Meaning.

This assertion requires a direct confrontation with physical ontology. Proponents of the Everettian Many-Worlds interpretation argue that probability is an illusion—a mere subjective artifact of self-location within a deterministic, branching multiverse of physical hardware. Conversely, strict Subjective Bayesians argue that probability does not exist at all outside the personal, betting preferences of a conscious agent.

The Geometric Theory of Ignorance rejects both the physical extravagance of Many-Worlds and the arbitrary solipsism of Subjectivism. We propose an objective "0-Worlds" ontology, rooted firmly in the Information Geometry of E.T. Jaynes [Ro3] and the Entropic Dynamics of Ariel Caticha [Cat12].

In the 0-Worlds paradigm, there is no underlying, hidden physical "stuff" that branches, nor are probabilities mere personal feelings. The probability distribution is the *objective, invariant geometric relation* between the data and the hypothesis space. If the laws of physics—including

continuous spacetime and particle mass—can be derived as the macroscopic limits of an optimal statistical inference engine resolving topological loops, then the universe requires no fundamental physical hardware. There are zero physical worlds. There is only the invariant geometry of information processing. A label without a probability is ontologically void; it is the objective, geometric assignment of probability that breathes meaning—and therefore, physical reality—into existence.

2 Why Entropy?

A surprisingly simple, asymptotic combinatorial argument, already shows the naturalness of the priors to come. Consider a large number n of draws, with replacement from a box of k types of tickets labeled $1, 2, \ldots, k$. Let us further assume that the dis- tribution of the different types of tickets in the box is $q = (q_1, q_2, \ldots, q_k)$, i.e., for $j = 1, 2, \ldots, k$ there is a proportion q_j of tickets with the label j in the box, with $0 < q_j < 1$ and $q_1 + q_2 + \cdots + q_k = 1$. If the tickets in the box have the same chance of being chosen, then the probability of observing a distribution $p = (p_1, p_2, \ldots, p_k)$ in the sample of n is given by,

$$W_n = \frac{1}{Z_n} \frac{e^{-nI(p/q)}}{\sqrt{p_1 p_2 \cdots p_k}} (1 + o(1))$$

as $n \to \infty$ Where $Z_n = (2\pi n)^{(k-1)/2}$ and, $I(p : q)$ denotes the standard Kullback number between the normalized distributions p and q. i.e.,

$$I(p : q) = \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}$$

The proof is a simple exercise in the use of the addition and multiplication rules of probability, together with Stirling's approximation:

$$\log n! = (n + 1/2) \log n - n + \log \sqrt{2\pi} + o(1)$$

Let $x_j = np_j$ be the observed number of tickets of type j in the sample of n. Without loss of generality we can assume the x_j to be integers. Otherwise, either replace by nearest integer or just consider p so that np_j are integers. Thus,

$$W_n = P(p|q, n) = P(x_1, \ldots, x_k|q, n) = \frac{n!}{x_1! x_2! \cdots x_k!} q_1^{x_1} \cdots q_k^{x_k}$$

is the standard multinomial probabilities. By the strong law of large numbers, the assumption that all the $q_j > 0$, implies that all the x_j will be arbitrarily large (with probability 1) as n approaches infinity. Thus, we are justified to use Stirling's approximation not just for n but for all the x_j as well. Thus, after a little simplification we get,

$$\log P(p|q, n) = -n \left[\sum_{j=1}^{k} p_j \log \frac{p_j}{q_j} - \frac{k-1}{2} \frac{\log n}{n} - k \frac{\log \sqrt{2\pi}}{n} \right] - \frac{1}{2} \sum_{j=1}^{k} \log p_j + o(1)$$

collecting the terms of order $o(1)$ (which do not contain p nor q) inside the square brackets we obtain,

$$\log P(p|q,n) = -nI(p:q) - \log\sqrt{p_1 p_2 \cdots p_k} - (k-1)\log\sqrt{2\pi n} + o(1)$$

hence, exponentiating both sides and noticing that $e^{o(1)} = 1 + o(1)$ we obtain the result.

It is therefore only natural to define a family of prior distributions on the simplex of discrete distributions on k labels by:

$$P(p \in A|q,\alpha) = \int_A \frac{1}{Z_\alpha} \exp(-\alpha I(p:q)) dV(p)$$

where A is a measurable subset of the k-simplex and $dV(p)$ is the informa- tion volume element on the simplex i.e., proportional to the square root of the determinant of Fisher information, which for discrete distributions is $dp/\sqrt{p_1 p_2 \cdots p_k}$ Moreover, $\alpha > 0$ represents the number of virtual (not necessarily integer) draws.

The noninformative priors in this paper provide a rigorous justification of the above as maximizers of an invariant notion of ignorance and generalize this simple result for, all meaningful measures of separation between unnormalized probability distributions and not just for the manifold of discrete distributions but for all smooth finite dimensional statistical models.

3 The Category of DataTheory (DataTh)

To make the extraction of meaning rigorous, we formalize the Data Theory space as a Category. By identifying the unique invariants of this category, we prevent the injection of arbitrary, subjective assumptions.

Let **DataTh** be the category where:

- **Objects** are statistical data-objects $(x, p) \in S$, where x is the observable label in a data space X, and p is the unobservable probability distribution that generated x.
- **Morphisms** are sufficient transformations $f : S \to S'$ mapping $(x, p) \to (x', p')$. A transformation is defined as sufficient if and only if it can be *stochastically inverted*: from (x', p') we can produce a new object (x'', p) where x'' is drawn from the exact same underlying unobservable distribution p as the original x.

Let **Geom** be the category of Riemannian manifolds. The Amari-Chentsov theorem establishes that the Fisher information metric—and the one-parameter family of α-connections (corresponding to our δ-geometry)—are the unique invariant geometric structures under sufficient statistics. Categorically, this means there exists a unique, covariant Functor $\mathcal{F}$: **DataTh** $\to$ **Geom** that maps probability spaces to Riemannian manifolds equipped with δ-connections, such that stochastically invertible transformations in the data space strictly map to geometric isometries and connection-preserving maps in the parameter space.

Consequently, any objective method for extracting meaning (inference) must be **equivariant** under this Functor. The "Maximum Honesty" δ-projection defined below satisfies this categorical imperative: doing inference on the raw data, and doing inference on the sufficient statistic, yield the exact same geometric posterior.

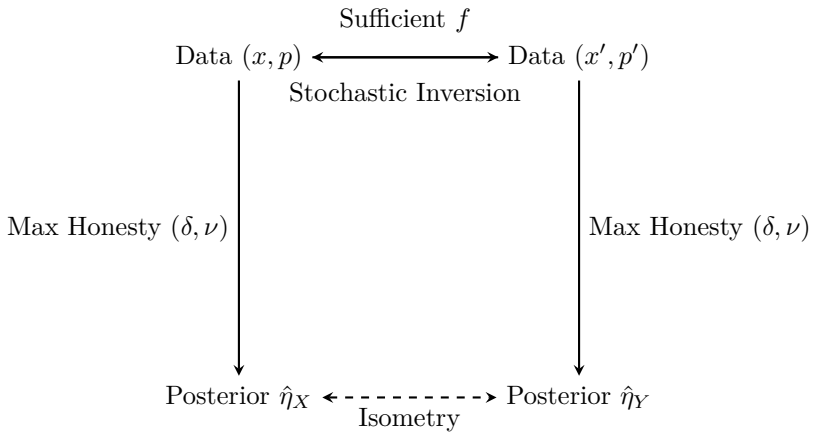

Figure 1: The Functorial Nature of Maximum Honesty. The δ-inference projection commutes with stochastically invertible transformations, ensuring the posterior is a strict categorical invariant rather than a subjective choice.

4 The Statistical No-Cloning Theorem

A profound consequence emerges from the formalization of the **DataTh** category. In quantum mechanics, the Wootters-Zurek No-Cloning Theorem (1982) dictates that an unknown quantum state cannot be perfectly duplicated, owing to the linearity of unitary evolution. Because our framework posits that "Probability is Meaning," and embeds distributions geometrically, an exact structural analogue must exist in DataTheory.

To "clone" an unknown data-object (x, p) means applying a purely observational process—a stochastic transformation or Markov kernel $K(x_1, x_2|x)$ independent of p—to produce two new labels x_1 and x_2 such that both are independently distributed according to the original, unknown p. That is, the required joint distribution must be the tensor product: $P(x_1, x_2) = p(x_1)p(x_2)$.

Theorem 1 (No-Cloning of Unknown Probabilities) *There exists no Markov kernel $K(x_1, x_2|x)$ independent of p that can perfectly clone an unknown data-object (x, p) into two independent data-objects (x_1, p) and (x_2, p) for all $p \in \mathcal{P}$, provided $\mathcal{P}$ contains at least two distinct distributions and their convex combinations.*

Proof. A generic stochastic transformation K acts as a linear operator

on the space of distributions:

$$K[p] = \int K(x_1, x_2|x)p(x)dx \tag{1}$$

For K to be a universal cloner, it must map any unknown input distribution p to the joint independent distribution $p \otimes p$. Let $p, q \in \mathcal{P}$ be distinct distributions. Due to the linearity of the Markov kernel, applying the cloner to a statistical mixture $\lambda p + (1-\lambda)q$ (where $\lambda \in (0,1)$) yields:

$$K[\lambda p + (1-\lambda)q] = \lambda K[p] + (1-\lambda)K[q] = \lambda(p \otimes p) + (1-\lambda)(q \otimes q) \tag{2}$$

However, the required target state for perfect cloning of this mixed distribution is the strictly non-linear tensor product:

$$(\lambda p+(1-\lambda)q)\otimes(\lambda p+(1-\lambda)q) = \lambda^2(p\otimes p)+(1-\lambda)^2(q\otimes q)+\lambda(1-\lambda)(p\otimes q+q\otimes p) \tag{3}$$

Equations (2) and (3) are only equal if $p = q$ or if the system is completely known ($\lambda \in \{0, 1\}$). ■

The classical intuition that one can simply "copy" the observed label by setting $x_1 = x$ and $x_2 = x$ (using a kernel composed of Dirac deltas $\delta_x(x_1)\delta_x(x_2)$) does not clone the meaning. It produces a joint distribution $P(x_1, x_2) = p(x_1)\delta_{x_1}(x_2)$, which represents absolute classical correlation (entanglement of the labels), not independent samples of the distribution.

You cannot clone meaning without destroying it or introducing correlated noise. The impossibility of perfectly amplifying an unknown statistical sample is the exact geometric dual of the quantum restriction, further reinforcing the 0-Worlds ontology: the limits of physical reality are the limits of information processing.

5 Fisher Information

Spaces of probability distributions are clearly not closed under addition and multiplication by scalar. They are not vector spaces. Nevertheless, there are canonical embeddings of probability distributions on Banach spaces that respect the **DataTh** Category.

For $0 < \delta < 1$ the Banach space $L_{1/\delta}$ of δ powers of measures contains the δ coordinates $l_\delta(p)$ of a probability distribution p with, $l_\delta(p) = p^\delta/\delta$. The Banach space associated to $1-\delta$ is the topological dual of the space associated to δ.

The only Hilbert space is the self-dual L_2 associated to $\delta = 0.5$. Fisher information is just the metric induced on the model as it is embedded into L_2 In other words: label the probabilities p in your model with the vector $2\sqrt{p}$ in the Hilbert space L_2 of square integrable functions. The Information metric at p is the matrix $g(p) = (g_{ij}(p))$ with components,

$$g_{ij}(p) = \int \partial_i(2\sqrt{p})\partial_j(2\sqrt{p})dx = \int (\partial_i \log p)(\partial_j \log p)p \; dx$$

where ∂_i is the partial derivative w.r.t. the i-th coordinate vector, and the integrals are over the space of x. Parametric statistical models with

smooth parametrizations are Riemannian manifolds with the information metric $g(p)$.

6 δ-separation, Entropy, and Duality

Let $l_0(p) = \log(p)$. For $0 \leq \delta \leq 1$ define the δ-separation between (possibly unnormalized) distinct distributions p and q by the positive number $I_\delta(p:q) = I_{1-\delta}(q:p)$ given for $0 < \delta < 1$ by:

$$I_\delta(p:q) = \frac{1}{\delta(1-\delta)} \int [\delta p + (1-\delta)q - p^\delta q^{1-\delta}] dx$$

and by the corresponding limit when $\delta \in \{0, 1\}$. Thus,

$$I_0(p:q) = \int \left(q - p + p \log \frac{p}{q} \right) dx = I_1(q:p).$$

The entries $g_{ij}(p)$ of the information matrix at p are also given by the duality product between the coordinates $l_\delta(p) \in L_{1/\delta}$ and the dual coordinates $l_{1-\delta}(p) \in L_{1/(1-\delta)}$ as,

$$g_{ij}(p) = \int (\partial_i l_\delta(p))(\partial_j l_{1-\delta}(p)) dx$$

The $I_\delta(p:q)$ numbers are invariants of the **DataTh**-category i.e., they are invariant under sufficient transformations. In fact, following the Amari-Chentsov characterization, the δ-separations (and their corresponding Amari α-connections) are the *only known* measures of information separation that preserve sufficiency—whether they are the absolutely unique invariant divergences for all general statistical models remains a profound open conjecture in information geometry.

7 Categorically Sound Objectives

Having observed data arbitrarily labeled $y_1, y_2, \ldots, y_n$ in a background of prior information how should we proceed extracting meaning to best predict unobserved y_{n+1}? What should we optimize? What should the target be?

In our current deep learning frenzy, Ed Jaynes's words resonate with renewed power: *Maximize Ignorance subject to whatever is assumed to be known!* This is just Maximum Honesty. An ethical principle.

7.1 The Actions of Ignorance

The risk functionals $\mathcal{A} = \mathcal{A}(t, \eta)$ defined for a Data Theory space in the **DataTh**-category, rank the pairs (t, η) according to their information, i.e., separation from ignorance. Here $t = t(y)$ is a (not necessarily normalized) distribution on the space of data labels y and $\eta = \eta(p)$ is a (not necessarily normalized) distribution on the hypothesis space M of possible theories p.

We interpret t as the (unknown) true distribution for the data and η as the (unknown) prior distribution on M. Let $\pi_0 = \pi_0(p)$ be a given fix prior distribution on the hypothesis space M of possible theories p.

This π_0 will be taken as a diffuse pre-prior on M. When the information volume $vol(M)$ is finite, we take $\pi_0(p) = 1$ as the uniform distribution on M. We always write distributions on the Riemannian manifold M as scalar density fields relative to the invariant Riemannian volume form dp on M.

Denote by $t\eta = P(y,p) = t(y)\eta(p)$. i.e. y and p are chosen independently. First pick $p \in M$ according to the distribution η, then independently choose label y in the data space according to the (true) distribution t.

Consider now any other distribution on (y,p), $P(y,p) = p(y)\eta(p) = p\eta$. Then, $I_\delta(p\eta : t\eta)$ measures the δ-separation between the joint distributions $p\eta$ and $t\eta$ and $I_{1-\nu}(\eta : \pi_0)$ measures the separation between the priors η and π_0.

For $\beta > 0$ define the positive scalar action,

$$\mathcal{A}(t,\eta) = \beta\nu I_\delta(p\eta : t\eta) + I_{1-\nu}(\eta : \pi_0) \tag{4}$$

The (unnormalized) pair (t,η) that minimizes the action $\mathcal{A}$ is,

$$\eta(p) = [1 + \beta\nu I_\delta(p : t)]^{-1/\nu}\pi_0(p) \tag{5}$$

$$t^\delta(y) = \frac{\int p^\delta(y)\eta(p)dp}{\int \eta(p)dp} \tag{6}$$

The action (4) does not contain derivatives and its optimization is a sim- ple problem in the calculus of variations. Just take derivatives equal to zero as if the functions were real variables. The expressions (5) and (6) pack a considerable amount of information in a short space.

Think of (5) as defining a kernel $k(p,t)$ on the dual Banach spaces associated to δ and $1-\delta$. It is a measure of separation between the probability distributions p and t on the data labels. The kernel (5) is literally the prior distribution of maximum ignorance.

The expression (6) provides the δ coordinates of t as the average of the δ coordinates of p or equivalently as the mean ker- nel. This is remarkably similar to the Reproducing Kernel Hilbert Space (RKHS) embeddings of probability distributions but in Banach spaces in- stead. The big difference with the RKHS approach is that here the kernel is fixed by the theory as the most ignorant prior given the choice of S. Besides, all the expressions are invariant under sufficient transformations preserving the **DataTh**-Category.

Notice that $t(y)$ is given by (6) as the length of a function measured in the Reproducing Kernel Banach Space associated to δ with kernel (5).

7.2 Maximum Likelihood and Maximum Honesty

The new way to statistical inference provided by Maximum Honesty is very simple. If data labels $y_1, y_2, \ldots, y_n$ are observed, just plugin the empirical distribution $\hat{t}_n$ instead of t in (5) and predict with (6).

Notice that the Maximum Likelihood Estimator (MLE) $\hat{p} \in M$ is the $\delta = 1$ projection of the empirical onto M.

$$\hat{p} = \arg\min_{p \in M} I_1(p : \hat{t}_n) \equiv \arg\min_{p \in M} \int \left[p - \hat{t}_n + \hat{t}_n \log \frac{\hat{t}_n}{p} \right] dy$$

$$= \arg\max_{p \in M} \left[\sum_{i=1}^{n} \log p(y_i) - \int p(y) dy \right]$$

where we have used the (unnormalized) empirical, $\hat{t}_n(y) = \sum_{i=1}^{n} \delta(y - y_i)$ and the fact that $I_1(p : \hat{t}_n) = I_0(\hat{t}_n : p)$.

8 Direct Posteriors and Statistical Learning Theory

When the unknown true distribution t is replaced by the unnormalized empirical $\hat{t}_n$ in (5) we get,

$$\hat{\eta}(p) = [1 + \beta\nu I_\delta(p : \hat{t}_n)]^{-1/\nu} \pi_0(p) \tag{7}$$

$$\hat{t}^\delta(y) = \int p^\delta(y) \hat{\eta}(p) dp \tag{8}$$

we call $\hat{\eta}(p)$ the (unnormalized) direct posterior with parameters β, ν, and δ. We call (8) the δ-coordinates of the δ-predictive distribution.

The special case when $\delta = 1$, $\nu \to 0$ and $\beta = 1$ is singled out as particularly important. By taking the limit when $\nu \to 0$ in (7) with $\delta = 1$ and $\beta = 1$, we get,

$$\hat{\eta}(p) = \exp(-nI_0(\hat{t}_n : p)) \pi_0(p) \tag{9}$$

If $\int p dy = 1$, i.e. for normalized probability distributions, (9) gives the unnormalized direct posterior:

$$\hat{\eta}(p) = p(y_1)p(y_2) \cdots p(y_n) \pi_0(p) \tag{10}$$

that we recognize as the unnormalized posterior distribution when the likelihood is $\prod_{i=1}^{n} p(y_i)$ and the prior is $\pi_0(p)$.

Moreover, with (10) and $\delta = 1$ replaced in (8) we get,

$$\hat{t}(y) = \int p(y) \left(\prod_{i=1}^{n} p(y_i) \right) \pi_0(p) dp \tag{11}$$

that we recognize as the standard Bayesian unnormalized predictive distri- bution. In other words with this special choice of parameters the inference is as if we had used Bayes Theorem but we did not! Bayesian Inference was produced automagically as a special case of maximum ignorance. Maximum honesty is more general than Bayesian Inference.

8.1 Categorical Derivation of SafeBayes and PAC-Bayes

The Direct Posterior (Equation 7) generated by maximizing categorical honesty yields a fractional power of the likelihood, governed by the parameter $\delta \leq 1$. In the modern statistical learning literature, this is known as a generalized Gibbs posterior, and its necessity has been heavily empirically documented.

Grünwald and van Ommen (2017) [Gru17] demonstrated through the SafeBayes framework that standard Bayesian inference ($\delta = 1$) is fundamentally inconsistent under model misspecification, requiring a "learning rate" $\eta < 1$ to temper the likelihood and restore convergence. Our parameter δ is exactly this learning rate. Furthermore, the Action of Ignorance (Equation 4) is structurally identical to the PAC-Bayesian minimization bounds pioneered by Catoni (2007) [Cat07] and McAllester, which balance empirical risk against the Kullback-Leibler divergence from the prior.

However, there is a profound philosophical and mathematical distinction. While SafeBayes and PAC-Bayes justify these fractional posteriors via algorithmic robustness, prequential coding, and empirical risk bounds, the Geometric Theory of Ignorance derives them from first principles. The δ-posteriors are the unique, invariant minimizers of the **DataTh** category. They are not ad-hoc engineering fixes for misspecified models; they are the fundamental, geometrically required solutions for processing information in Banach spaces of unnormalized distributions.

9 Information Metric of Logistic Regression

The Fisher information matrix is given at $p = P_{w,x}$ by,

$$g(w) = E_w\{(\nabla \log p(y|x,w))^T (\nabla \log p(y|x,w))\}$$

where E_w denotes expectation w.r.t. $P_{w,x}$ and ∇ is the gradient w.r.t. w. Let's denote by $l = l(w) = \log p(y|x,w)$. Then,

$$l(w) = \sum_{i=1}^{n} \{y_i \log \theta_i + (1-y_i)\log(1-\theta_i)\}$$

therefore,

$$\nabla l(w) = \sum_i \left\{ \frac{y_i}{\theta_i} - \frac{1-y_i}{1-\theta_i} \right\} \nabla \theta_i$$

and,

$$\nabla \theta_i = \theta_i(1-\theta_i)x_i = \frac{x_i}{2(1+\cosh(x_i w))}$$

substituting the above into the expression for $\nabla l(w)$, we obtain $\nabla l(w) = \sum_i (y_i - \theta_i)x_i$. Hence,

$$(\nabla l)^T(\nabla l) = \sum_{i,j}(y_i - \theta_i)(y_j - \theta_j)x_i^T x_j$$

therefore, taking expectations with the probability distribution $P_{x,w}$ and recalling that the y_i are independent for different values of i we get,

$$g(w) = \sum_{i,j} cov(y_i, y_j)x_i^T x_j = \sum_i \theta_i(1-\theta_i)x_i^T x_i$$

thus, we finally obtain two useful expressions for the information metric,

$$g(w) = \frac{1}{2}\sum_{i=1}^{n} \frac{x_i^T x_i}{1+\cosh(x_i w)} \tag{12}$$

which in matrix form is,

$$g(w) = \frac{1}{2}x^T diag\left(\frac{1}{1+\cosh(xw)}\right)x \tag{13}$$

9.1 The Information Volume of Logistic Regression

The hypothesis space M generated by the logistic regression model is a Riemannian manifold of dimension k with metric tensor given by the information matrix at each w by the expression (13). The volume form on a Riemannian manifold with metric $g(w)$ is given in the $w = (w^j)$ coordinates by,

$$dp = dV(w) = \sqrt{\det g(w)}dw^1 \wedge dw^2 \cdots \wedge dw^k$$

with total (k-dim) volume given by integrating the volume element over M:

$$vol(M) = \int_M dV = \int_{R^k} \sqrt{\det g(w)}dw.$$

When x is of full rank so that $\det x^T x > 0$, we have $vol(M) < \infty$. When $k = n$, i.e. when x is a square k by k matrix, the computation of the determinant is trivial using expression (13). Thus,

$$\det g(w) = 2^{-k}\det(x^T x)\prod_{j=1}^{k}\frac{1}{1+\cosh(x_j w)}$$

$$vol(M) = 2^{-k/2}|\det x|\int \frac{dw}{\sqrt{\prod_{j=1}^{k}(1+\cosh(x_j w))}}$$

$$= 2^{-k/2}\int \frac{du}{\sqrt{\prod_{j=1}^{k}(1+\cosh(u_j))}}$$

$$= 2^{-k/2} \left(\int_{-\infty}^{\infty} \frac{dz}{\sqrt{1+\cosh(z)}} \right)^k$$
$$= 2^{-k/2} (\sqrt{2}\pi)^k = \pi^k$$

where we performed the linear change of variables $u = xw$.

Theorem 2 *When x is a square of full rank the information volume of the logistic regression model is independent of x and has value π^k where k is the dimension of the manifold.*

9.2 Curvature in the Exactly Parameterized Case

Recall that the components Γ_{ijk} of the Levi-Civita metric connection are,

$$2\Gamma_{ijk} = \partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}.$$

With the help of SageMath symbolic computation for the exactly parameterized case ($n = k = 2$):

```
def g_logistic_metric(k,n):
    assert n >= k, "n=%d must be at least k=%d" % (n,k)
    for j in range(k): var('w%d'%j)
    w = [var('w%d'%j) for j in range(k)]
    x = matrix(SR,n,k)
    for i in range(n):
        for j in range(k):
            x[i,j] = var('x%d%d'%(i,j))
    # ... (Metric construction omitted for brevity) ...
    return g

g = g_logistic_metric(2,2)
Sim(g.R)
```

It simplifies to zero!

For the exactly parameterized case ($n = k = 2$), SageMath symbolic computation reveals the Ricci scalar vanishes ($R = 0$). However, this does not imply that the general $n \times k$ logistic regression manifold (where $n \gg k$) is flat. The curvature for the overparameterized regime remains an open question, and its geometric complexity likely requires the full integration of the boundary conditions where some $\theta_i \in \{0, 1\}$.

References

[Am] Amari, S.-i.: Differential-Geometrical Methods in Statistics. (Lecture Notes in Statistics, Vol. 28). Springer-Verlag 1985

[Cat07] Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. IMS Lecture Notes Monograph Series, Vol. 56 (2007)

[Cat12] Caticha, A.: Entropic Inference and the Foundations of Physics. EBEB 2012, São Paulo, Brazil (2012)

[Gru17] Grünwald, P., van Ommen, T.: Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. Bayesian Analysis 12(4), 1069-1103 (2017)

[I-H] Ibragimov, I., Has'minskii, R.: Statistical estimation. (Applications of Mathematics, Vol. 16). Springer-Verlag 1981

[Je] Jeffreys, H.: Theory of Probability. Oxford University Press 1939

[Ku] Kullback, S.: Information Theory and Statistics. John Wiley, New York 1959

[L-S] Lai, T., Siegmund, D. (eds.): Herbert Robbins Selected Papers. Springer-Verlag 1985

[Ro1] Rodríguez, C.: The metrics induced by the kullback number. In: Skilling, J. (ed.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers 1989

[Ro2] Rodríguez, C.: Objective bayesianism and geometry. In: Fougere, P. F. F. (ed.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers 1990

[Ro3] Rosenkrantz, R. (ed.): E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics. (Vol. 158). Synthese Library 1983