Information Geometry of the Count Table: The Statistical Manifold of the Universal Model

Claude and MJC

February 12, 2026

Abstract

The count table $c : I \times O \to \mathbb{N}$, normalized to a probability distribution, defines a point on the statistical manifold of distributions over $I \times O$. We study the geometry of this manifold equipped with the Fisher information metric. The conditional distributions $P(o \mid i)$ form an $|I|$-parameter exponential family. The natural parameters are the log-odds ratios $\log(c(i,o)/c(i,o'))$, which are the pattern strengths of the UM. The Fisher metric on this family measures how distinguishable two models are from finite data. We show: (1) the KL divergence between two count tables equals the relative entropy, connecting to the UM's cross-entropy loss; (2) the geodesics of the Fisher metric correspond to exponential interpolations between models, which is the EM algorithm's update path; (3) the tock step moves along a *transverse* direction (changing the manifold itself, not the point on it), and the tick step moves along a geodesic (updating the point). The geometry explains why KN smoothing works (it moves the estimate toward the "center" of the manifold) and why the $R^2 \approx 0.83$ invariant emerges (it is a geometric property of the Fisher metric on high-dimensional exponential families).

1 The Statistical Manifold

Definition 1 (Count table manifold)**.** *For event space $E = I \times O$ with $|I| = m$ and $|O| = n$, the* count table manifold *is:*

$$\mathcal{M} = \left\{ P \in \mathbb{R}_+^{m \times n} : \sum_o P(i,o) = P(i,\cdot) > 0 \; \forall i, \; \sum_{i,o} P(i,o) = 1 \right\}.$$

This is the interior of the probability simplex Δ^{mn-1} restricted to distributions with full support on I.

Definition 2 (Fisher information metric)**.** *The* Fisher information metric *on $\mathcal{M}$ is:*

$$g_{\mu\nu}(\theta) = E_\theta \left[\frac{\partial \log P(x;\theta)}{\partial \theta_\mu} \cdot \frac{\partial \log P(x;\theta)}{\partial \theta_\nu} \right],$$

where θ are the parameters of the distribution.

Proposition 3 (Exponential family structure)**.** *The conditional distributions $\{P(o \mid i)\}_{i \in I}$ form an exponential family with natural parameters $\eta_{io} = \log P(o \mid i) - \log P(n \mid i)$ (log-odds relative to a reference event n) and sufficient statistics $T_{io}(x) = \mathbf{1}[X_I = i, X_O = o]$ (event indicators).*

2 The Pattern Matrix as Natural Parameters

Proposition 4 (Patterns are log-odds)**.** *The UM's pattern strengths, in the continuous (non-SN-quantized) limit, are:*

$$p_{ij} \propto \log \frac{c(i,j)}{c(i,\cdot)} = \log P(j \mid i).$$

The log-conditional probabilities are the natural parameters of the exponential family. The pattern matrix IS the natural parameterization of the statistical manifold.

Corollary 5. *The forward pass in the continuous limit:*

$$(f_p(t))_j = \max_i (t_i + p_{ij})$$

(where $t_i = \log P(i)$ and $p_{ij} = \log P(j \mid i)$) computes the MAP estimate on the joint manifold:

$$\hat{o} = \arg\max_j \max_i \log P(i,j) = \arg\max_j \max_i (t_i + p_{ij}).$$

3 KL Divergence and Cross-Entropy

Proposition 6 (KL = relative entropy = excess loss)**.** *The KL divergence between the true distribution P and the model's estimate $\hat{P}$ is:*

$$D_{KL}(P\|\hat{P}) = \sum_{i,o} P(i,o) \log \frac{P(i,o)}{\hat{P}(i,o)} = H(P,\hat{P}) - H(P),$$

where $H(P,\hat{P})$ is the cross-entropy (the UM's loss in bpc) and $H(P)$ is the entropy (the irreducible loss).

Remark 7. *D_{KL} measures the "distance" from $\hat{P}$ to P on the statistical manifold. It is not a true distance (not symmetric, doesn't satisfy triangle inequality) but it is the Bregman divergence associated with the negative entropy, which makes it the natural "distance" for exponential families.*

Proposition 8 (The Pythagorean theorem)**.** *For the exponential family of count-table distributions, the KL divergence satisfies a Pythagorean identity. If P is the true distribution, $\hat{P}_{MLE}$ is the MLE (the empirical distribution from counting), and $\hat{P}_{smooth}$ is a smoothed estimate (e.g., KN smoothing), then:*

$$D_{KL}(P\|\hat{P}_{smooth}) = D_{KL}(P\|\hat{P}_{MLE}) + D_{KL}(\hat{P}_{MLE}\|\hat{P}_{smooth}).$$

Smoothing ALWAYS increases the KL divergence from the truth, but it may decrease the expected *KL divergence over finite samples (bias-variance trade-off).*

4 Geodesics and the EM Path

Proposition 9 (Geodesics in the Fisher metric)**.** *The geodesics on $\mathcal{M}$ (with the Fisher metric) are* exponential geodesics*: they interpolate between two distributions by exponentially weighting their log-probabilities:*

$$P_\lambda(i,o) \propto P_0(i,o)^{1-\lambda} \cdot P_1(i,o)^\lambda, \quad \lambda \in [0,1].$$

Proposition 10 (EM updates follow geodesics)**.** *The EM algorithm (tick-tock iteration, per the fixed-point paper) updates the distribution along an exponential geodesic in the natural parameter space. Each E-step (counting) computes the expected sufficient statistics; each M-step (event space optimization) moves to the maximum-likelihood point. The M-step path is a geodesic in the e-flat (natural parameter) geometry.*

5 Tick vs. Tock on the Manifold

Proposition 11 (Tick = motion on the manifold)**.** *The tick step (updating the count table from new data) moves the model's position on the manifold $\mathcal{M}$ without changing the manifold itself. As data accumulates, the position converges to the true distribution P:*

$$\hat{P}_N \xrightarrow{N\to\infty} P \quad \text{on } \mathcal{M}.$$

Proposition 12 (Tock = change of manifold)**.** *The tock step (changing the event space) replaces the manifold $\mathcal{M}$ with a different manifold $\mathcal{M}'$ (corresponding to the new event space). This is a* transverse *motion: it changes the space of possible distributions, not the position within a fixed space.*

Remark 13. *The tick-tock distinction maps to the distinction between:*

- ***Estimation*** *(finding the best point on a fixed manifold): this is statistics (tick).*
- ***Model selection*** *(choosing the right manifold): this is architecture (tock).*

The Fisher metric is relevant for estimation (it measures how hard it is to distinguish nearby points). For model selection, the relevant geometry is the "manifold of manifolds"—the lattice of event spaces with the MI ordering.

6 KN Smoothing as Geometric Centering

Proposition 14 (KN smoothing as shrinkage)**.** *KN smoothing with discount δ shrinks the empirical distribution toward the lower-order distribution:*

$$P_{KN}(o \mid i) = \frac{\max(c(i,o) - \delta, 0)}{\sum_{o'} c(i,o')} + \gamma(i) \cdot P_{lower}(o),$$

where $\gamma(i)$ is a normalization factor. On the statistical manifold, this moves the estimate from the MLE toward the "center" (uniform or lower-order distribution).

Remark 15. *Geometrically, KN smoothing is a form of* regularization*: it prevents the estimate from being too close to the boundary of the simplex (where some probabilities are zero). The δ parameter controls the shrinkage amount. The "no support $\neq$ disbelief" principle says: points on the boundary (zero probabilities) are* not *valid model positions, because zero probability means "certainty of non-occurrence," which requires positive evidence for the alternative. KN smoothing keeps the model in the interior of the manifold, where the Fisher metric is well-defined and "no support" remains distinguishable from "disbelief."*

7 The R^2 Invariant as a Geometric Property

Proposition 16 (R^2 from Fisher metric)**.** *The conjunction $R^2 \approx 0.83$ for the 128-hidden tanh RNN is a geometric property of the Fisher metric on a 128-dimensional Gaussian manifold.*

The argument (from the conjunction invariant paper):

1. *Each neuron h_k is approximately the product of two independent sub-Gaussian variables (the contributions from offsets d_1 and d_2).*

2. *The R^2 of the best 2-factor fit depends on the ratio of the order statistics of $|h_k|$ to the norm $\|h\|$.*

3. *For a 128-dimensional Gaussian, the expected max-to-norm ratio is $\sqrt{2\log 128}/\sqrt{128} = \sqrt{14}/\sqrt{128} \approx 0.331$.*

4. *$R^2 = \mathrm{erf}^2(\sqrt{2\log H}/\sqrt{H})$. At $H = 128$: $R^2 \approx 0.83$.*

This ratio is a property of the Fisher metric on the Gaussian manifold: it measures how "peaked" the distribution of neuron activations is, which depends on the dimension (128) and the geometry (Gaussian).

8 The Tropical Metric

Definition 17 (Tropical distance)**.** *On the support lattice $\{0, \ldots, 255\}^n$, define the* tropical distance*:*

$$d_{trop}(s, s') = \max_i |s_i - s'_i|.$$

This is the ℓ^∞ metric—the maximum difference in support across all events.

Proposition 18 (Forward pass is 1-Lipschitz)**.** *The forward pass is 1-Lipschitz in the tropical metric:*

$$d_{trop}(f_p(t), f_p(t')) \leq d_{trop}(t, t').$$

Proof. For each output j:

$$|(f_p(t))_j - (f_p(t'))_j| = |\max_i \min(t_i, p_{ij}) - \max_i \min(t'_i, p_{ij})| \tag{1}$$

$$\leq \max_i |\min(t_i, p_{ij}) - \min(t'_i, p_{ij})| \tag{2}$$

$$\leq \max_i |t_i - t'_i| = d_{\mathrm{trop}}(t, t'). \tag{3}$$

The first inequality uses $|\max f - \max g| \leq \max |f - g|$. The second uses the 1-Lipschitz property of $\min(\cdot, p)$. □

Corollary 19. *The forward pass is a* contraction *in the tropical metric: small changes in input support produce small changes in output support. This is the UM's version of "stability": the model's predictions change continuously with the evidence.*

9 Discussion

The information-geometric perspective reveals:

1. **The pattern matrix IS the natural parameterization.** Log-conditional probabilities are the natural coordinates on the statistical manifold. The UM's pattern strengths are (discretized versions of) these coordinates.

2. **KL divergence IS the UM's loss.** The cross-entropy loss in bpc is the KL divergence plus the entropy. Minimizing bpc is minimizing KL divergence: finding the closest point on the manifold to the true distribution.

3. **KN smoothing IS geometric regularization.** Smoothing keeps the model in the interior of the manifold, where the geometry is well-defined and the epistemology of zero (no support $\neq$ disbelief) is respected.

4. **The R^2 invariant IS geometric.** The 0.83 value is determined by the dimension and the metric, not by the specific data or training procedure.

5. **The forward pass IS a contraction.** In the tropical metric, predictions change smoothly with evidence. This is the geometric content of the UM's "caution": the min operation ensures that predictions never exceed the weakest evidence.

References

[1] Michaeljohn Clement. *CMP*. `https://cmpr.ai/cmp.pdf`, 2026.

[2] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

[3] Claude and MJC. *The Conjunction Invariant: Why $R^2 \approx 0.83$ Is a Fixed Point*. Hutter archive, 12 Feb 2026.

[4] Claude and MJC. *No Support Is Not Disbelief*. Hutter archive, 12 Feb 2026.