

The KL Divergence Is Not a Distance, but Its Second Derivative Is a Metric

— The Asymmetry Is Not a Defect but a Division of Labour: Every f-Divergence Collapses onto the Same Metric — [Paper 137]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
August 2026

— *From thinking to syncing.* —

ABSTRACT

The Kullback–Leibler divergence, the relative entropy, is described as a measure of how far apart two distributions are. **This paper writes that it is not a distance, that it nevertheless is one infinitesimally, and that its asymmetry is doing work. NO NEW MATHEMATICAL THEOREM AND NO NEW PHYSICAL LAW IS CLAIMED. First, it is not a distance.** It is asymmetric (**0.223805** against **0.193794**), and **the triangle inequality fails**: over **200,000** random triples it failed in **25.4%** of cases (Section 2). **Second, one of the two cores. Its second derivative is a metric.** Shifting the mean of a Gaussian family by ϵ gives $D = \epsilon^2/(2\sigma^2)$, which agrees exactly with half the Fisher metric, $\frac{1}{2}g_{\mu\mu}\epsilon^2$. Shifting the width, D vanishes to first order and **its second derivative is exactly $g_{\sigma\sigma} = 2/\sigma^2$** (Paper 133). **Not a distance globally; a distance infinitesimally** (Section 3). **Third, the other core. Every f-divergence collapses onto the same metric.** Expanding with $p = a$ and $q = a + e$, KL gives $\frac{1}{2a}e^2$, reverse KL $\frac{1}{2a}e^2$, chi-squared $\frac{1}{a}e^2$ and squared Hellinger $\frac{1}{4a}e^2$: **all the same quadratic form up to a constant. That is what makes Chentsov's uniqueness possible at all** (Section 4). **Fourth, the asymmetry is not a defect but a division of labour.** Approximating an equal mixture of Gaussians at **−3** and **+6** by a single Gaussian, **the forward direction $D(\text{target}||\text{model})$ gives mean 1.50 and width 4.60, spreading to cover both modes**, while **the reverse direction $D(\text{model}||\text{target})$ gives mean −3.00 and width 1.00, collapsing onto one** (computed here). **The two directions answer different questions** (Section 5). **Fifth, in thermodynamics it is the excess free energy.** The free energy of a non-equilibrium distribution exceeds the equilibrium value by exactly $k_B T D(p||p_{\text{eq}})$. **Information and free energy are the same quantity in different units — the units are a convention** (Paper 131); **the non-negativity is not** (Section 6).

Scope of this paper: NO NEW MATHEMATICAL THEOREM AND NO NEW PHYSICAL LAW IS CLAIMED.

Kullback–Leibler (1951), Csiszár's f-divergences, Chentsov's theorem and the difference between forward and reverse KL in variational inference are **all standard**. What this paper adds is only **counting the failures of the triangle inequality, showing in closed form that the second derivative is the Fisher metric, setting the quadratic coefficients of four divergences side by side, and demonstrating the asymmetry as a division of labour**. The KL divergence is not called a distance.

1. Notation and position

For two distributions p, q write

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \quad (1)$$

(an integral in the continuous case). $D \geq 0$, with equality only when $p = q$.

Position: existing results are laid side by side and their reach written down. No new theorem is claimed. Paper 17 said that **only writing the addresses makes a claim checkable**. This paper does that for the word “apartness”.

2. It is not a distance

Verification 2.1 (asymmetry): For $p = (0.5, 0.3, 0.2)$ and $q = (0.2, 0.5, 0.3)$ this paper obtains

$$D(p||q) = 0.223805, \quad D(q||p) = 0.193794.$$

The same two distributions give different values when the order is reversed.

Verification 2.2 (the triangle inequality): Inserting $r = (0.34, 0.33, 0.33)$,

$$D(p||q) = 0.223805 > D(p||r) + D(r||q) = 0.064083 + 0.074746 = 0.138829,$$

so **the triangle inequality does not hold**. Over 200,000 random triples it failed in 50,826 cases (25.4%). (QED)

Observation 2.3 (it is still a measure of apartness)

$D \geq 0$ and it vanishes only when $p = q$, so whether two distributions are the same can be told. What cannot be told is a consistent ordering of which is further from which. One word, “distance”, contains two demands — deciding identity, and closing a triangle. KL meets only the first.

3. But its second derivative is a metric

Verification 3.1 (shifting the mean): Shifting only the mean of a Gaussian family by ϵ , the closed form gives

$$D(N(\mu, \sigma) \parallel N(\mu + \epsilon, \sigma)) = \frac{\epsilon^2}{2\sigma^2}.$$

The Fisher metric of Paper 133 has $g_{\mu\mu} = 1/\sigma^2$, so the difference from $\frac{1}{2}g_{\mu\mu}\epsilon^2$ is **zero** — **an exact agreement, not an approximation** (verified here). (QED)

Verification 3.2 (shifting the width): Shifting only the width by ϵ gives

$$D = \frac{\epsilon^2}{\sigma^2} - \frac{5\epsilon^3}{3\sigma^3} + O(\epsilon^4).$$

This paper confirms that at $\epsilon = 0$ **both the value and the first derivative vanish** and **the second derivative is exactly $2/\sigma^2 = g_{\sigma\sigma}$** . (QED)

Observation 3.3 (and the third-order term is what the asymmetry is)

$-\frac{5\epsilon^3}{3\sigma^3}$ changes sign with ϵ . Up to second order the divergence is symmetric; from third order it is not. That a thing can fail to be a distance globally while being one infinitesimally is exactly this term's doing.

Observation 3.4 (its place in the corpus)

Paper 133 wrote that the Fisher metric is **positive definite** and **unique in Chentsov's sense**. This paper writes where that metric comes from: it is the second derivative of a divergence.

4. Every f-divergence collapses onto the same metric

Verification 4.1: The contribution from one outcome whose probability moves from $p = a$ to $q = a + e$ was expanded.

divergence	quadratic term
KL	$\frac{1}{2a}e^2$
reverse KL	$\frac{1}{2a}e^2$
chi-squared	$\frac{1}{a}e^2$
Hellinger (squared)	$\frac{1}{4a}e^2$

All four are constants times e^2/a . (QED)

Observation 4.2 (this is what uniqueness rests on)

Four quantities that differ entirely at global scale collapse onto the same quadratic form infinitesimally. That is why “the invariant metric is the Fisher metric and no other” (Chentsov, Paper 133) can hold. Uniqueness is another way of saying that there was only ever one candidate.

Observation 4.3 (put the other way)

The quadratic coefficient does not tell which divergence one is looking at, since only a constant factor distinguishes them. The reason to prefer one divergence lies entirely outside second order — in how the tails are weighted and in what shape the optimum takes (Section 5).

5. The asymmetry is a division of labour

Verification 5.1 (one target, two answers): Taking an equal mixture of Gaussians at -3 and $+6$, each of width 1 , as the target, it was approximated by a **single** Gaussian.

what is minimised	mean	width	the D attained
$D(\text{target} \text{model})$ (forward)	1.50	4.60	0.8350
$D(\text{model} \text{target})$ (reverse)	-3.00	1.00	0.6931

The forward direction stands at the midpoint of the two modes and spreads to cover them; the reverse direction sits exactly on one of them. (QED)

Observation 5.2 (why)

In $D(p||q) = \sum p \log(p/q)$, a heavy penalty falls where p is large and q is small. In the forward direction p is the target, so the model is not allowed to vanish where the target lives, and it therefore covers. In the reverse direction p is the model, so the model is not allowed to live where the target vanishes, and it therefore dives into one mode.

Observation 5.3 (not a defect)

Reading the asymmetry as a defect to be repaired by symmetrisation is a mistake. The two directions answer different questions, and which to use is decided by what one wants. Symmetrising erases that distinction. As Paper 95 said, being a convention is not the same as being a free choice.

6. In thermodynamics it is the excess free energy

Observation 6.1

With the equilibrium distribution $p_i^{\text{eq}} = e^{-E_i/k_B T} / Z$ and the equilibrium free energy $F = -k_B T \log Z$, for any distribution p

$$\left(\sum_i p_i E_i - k_B T S[p] / k_B \right) - F = k_B T D(p||p^{\text{eq}}).$$

The free energy of a non-equilibrium distribution exceeds the equilibrium value by exactly $k_B T$ times the relative entropy.

Observation 6.2 (the relation to the second law)

$D \geq 0$ states that the free energy cannot fall below its equilibrium value, and that D decreases under relaxation is one way of stating the second law (Papers 88, 136).

Observation 6.3 (on units)

Information and free energy are the same quantity in different units. Whether to multiply by $k_B T$, and whether the logarithm is base 2 or base e , are both conventions (Paper 131). What is not a convention is the single fact $D \geq 0$.

7. Conclusion

1. The KL divergence is not a distance (Ver. 2.1, 2.2). It is asymmetric, and the triangle inequality fails for 25.4% of random triples. One word contains two demands, and KL meets only the first.
2. But its second derivative is a metric (Ver. 3.1, 3.2). For a shift of the mean it agrees exactly with $\frac{1}{2} g_{\mu\mu} \epsilon^2$; for a shift of the width the second derivative is exactly $g_{\sigma\sigma}$ (Paper 133).
3. The asymmetry is made by the third-order term (Obs. 3.3). $-\frac{5\epsilon^3}{3\sigma^3}$ changes sign. Symmetric to second order, asymmetric from third.
4. Every f-divergence collapses onto the same metric (Ver. 4.1): $\frac{1}{2a}, \frac{1}{2a}, \frac{1}{a}, \frac{1}{4a}$ — differing only by a constant. This is what Chentsov's uniqueness rests on.
5. The asymmetry is a division of labour (Ver. 5.1). Forward gives mean 1.50, width 4.60, covering; reverse gives mean -3.00, width 1.00, one mode. Symmetrising erases the distinction.
6. In thermodynamics it is the excess free energy (Obs. 6.1). The units are a convention; the non-negativity is not (Papers 131, 136).

Closing: The word “apartness” lumps together what can be measured and what cannot. What KL measures is whether two distributions are the same, not a consistent ordering of which is further. And shrink to the infinitesimal and every divergence becomes the same single metric — the differences live only where the shrinking stopped.

References

1. Kullback, S. & Leibler, R. A. (1951). *On information and sufficiency*. Annals of Mathematical Statistics, 22.
2. Csiszár, I. (1967). *Information-type measures of difference of probability distributions and indirect observations*. Studia Scientiarum Mathematicarum Hungarica, 2.
3. Chentsov, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nauka (English translation: American Mathematical Society, 1982).
4. Amari, S. & Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society.
5. Minka, T. (2005). *Divergence measures and message passing*. Microsoft Research Technical Report.

6. Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley.
7. Yamagishi, Y. (2026). *The Equals Sign Has Distinct Roots Too* [Paper 16], 3rd ed. Zenodo. 10.5281/zenodo.21600471
8. Yamagishi, Y. (2026). *The 8π Is an Artefact of the Units* [Paper 17], 3rd ed. Zenodo. 10.5281/zenodo.21600813
9. Yamagishi, Y. (2026). *Two Entropies* [Paper 88]. Zenodo. 10.5281/zenodo.21747809
10. Yamagishi, Y. (2026). *“Dimension” Is Not One Thing* [Paper 94]. Zenodo. 10.5281/zenodo.21758581
11. Yamagishi, Y. (2026). *Convention and Fact* [Paper 95]. Zenodo. 10.5281/zenodo.21758621
12. Yamagishi, Y. (2026). *The Past Hypothesis* [Paper 98]. Zenodo. 10.5281/zenodo.21758739
13. Yamagishi, Y. (2026). *The Crossing of Royal Roads* [Paper 100]. Zenodo. 10.5281/zenodo.21758864
14. Yamagishi, Y. (2026). *Extracting Signal from Noise* [Paper 106]. Zenodo. 10.5281/zenodo.21836984
15. Yamagishi, Y. (2026). *Scale Invariance Fixes an Exponent — in Four Different Ways* [Paper 117]. Zenodo. 10.5281/zenodo.22005515
16. Yamagishi, Y. (2026). *Of $k_B T \ln 2$, Only the T Is Physics* [Paper 131]. Zenodo. 10.5281/zenodo.22080240
17. Yamagishi, Y. (2026). *The 2 in Itô's Lemma Is Not a Dimension* [Paper 132]. Zenodo. 10.5281/zenodo.22080358
18. Yamagishi, Y. (2026). *The Fisher Metric Is Unique Only on the Classical Side* [Paper 133]. Zenodo. 10.5281/zenodo.22080550
19. Yamagishi, Y. (2026). *The Equality Is Exact, and the Average Is Held by the Tail* [Paper 136]. Zenodo. 10.5281/zenodo.22081230

— *From thinking to syncing.* —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)