# Yang Hu · Proposal day 1 · Wed 21

Area: Singular learning theory / training dynamics. Finite-temperature estimation of local learning coefficients against higher-order loss geometry.

Group: Yang Hu

## The question (about 40 words)

In a tanh network at an exact realisation with one sextic transverse direction, there is a declared range of inverse temperature and localisation, within a stated budget, where the exact localised coefficient exceeds the Hessian prediction by a persistent 1/6 absent in a regular nonlinear control, and SGLD resolves it.

**Terms (not counted).** Three quantities are kept apart: λ, the asymptotic learning coefficient (Watanabe's real log canonical threshold [7]); Λ(t, γ) = −t ∂_t log Z(t, γ) = t E[K] under the posterior ∝ exp(−tK − γ‖w − w*‖²/2) on a declared domain, the exact finite-temperature localised observable; and Λ̂, what the sampler returns. The Hessian prediction is G(t, γ) = ½ Σ_i th_i/(th_i + γ) over the Hessian eigenvalues h_i, exact for a quadratic loss. The measured excess decomposes as Λ̂ − G = (λ − r/2) + [(Λ − λ) − (G − r/2)] + (Λ̂ − Λ): the asymptotic geometric excess, the finite-temperature and localisation correction, and the sampling error. "Persistent" means the first term, the part that does not decay as t grows. At smooth points of the zero-loss set it is Kurumadani's (c − r)/(2k) under his nondegeneracy conditions [1, 2], with r the Fisher rank, c the codimension and k the transverse jet order; the sextic case is k = 3, excess 1/6.

## Why it matters (about 120 words)

Developmental interpretability reads local learning coefficient estimates along training as effective parameter counts, and Timaeus's own guide says the sampler's temperature, localisation and step size have no defaults and must be swept [6]. At smooth points of the zero-loss set, what the coefficient adds to curvature is a definite number with a definite cause, (c − r)/(2k). But a finite-temperature excess over the Hessian prediction is not evidence of it: the regular loss tanh²w gives Λ − G = 1/(2t), a fixed Gaussian localiser lets a distant flatter zero take over as t grows, and a nearly singular εv² term imitates the singular case for γ² ≪ t ≪ ε⁻². Whether a practical measurement can see higher-order geometry at all, in which range and at what cost, is what anyone reading an LLC curve needs to know. A negative result names the constraint that blocks detection.

## Why it is open (about 200 words)

Kurumadani [1] proves λ = r/2 + (c − r)/(2k) at smooth points of the realisable set under jet nondegeneracy, and his dissertation [2] computes the two-unit tanh case with zero input weights explicitly, $(d_1, r, m) = (2, 1, 3)$, λ = 2/3 under $a_1^2 \neq a_2^2$, with the four-unit extension; the asymptotic answer for the test case is settled, which is what makes it usable as ground truth. Wang et al.

[3], Appendix D.3, state λ ≥ r/2 and find Hessian-rank estimates unreliable. Cai [4] derives continuous-time Gaussian-localised targets for $u^2$ and $u^2v^2$ (a Bessel-function closed form for the centred product singularity), separates them from finite-step Langevin effects where a closed form exists, and gives the mean energy under a Gaussian surrogate (his Proposition 1), all in service of a thermodynamic-variance estimator on MLP trajectories; he has no regular nonlinear control, no smooth-fibre sextic geometry and no network prediction. Hitchcock and Hoogland [5] benchmark sampler fidelity to the local posterior. Nobody has written down a diagnostic that separates a persistent excess from the O(1/t) correction of a regular nonlinear minimum, derived the visibility scale t ≍ γ^k/a with the coefficient and localisation metric taken in a network's own coordinates, or tested whether an actual tanh network's 1/6 can be resolved within a budget. That is the gap.

# References


[1] Y. Kurumadani. Real Log Canonical Thresholds at Non-singular Points. arXiv:2408.13030 (Main Theorem 2 and Remark 2.5).

[2] Y. Kurumadani. Learning Coefficients in Semiregular Models. Doctoral dissertation, The University of Osaka, March 2026 (Examples 5.1(3) and 5.2).

[3] G. Wang, J. Hoogland, S. van Wingerden, Z. Furman, D. Murfet. Differentiation and Specialization of Attention Heads via the Refined Local Learning Coefficient. arXiv:2410.02984 (Appendix D.3).

[4] Y. Cai. Noise-Debiased Thermodynamic Variance for Local Learning Coefficient Probes. arXiv:2606.22389 (Section 4.1.1, Appendices A.2, B and D).

[5] R. Hitchcock, J. Hoogland. From Global to Local: A Scalable Benchmark for Local Posterior Sampling. arXiv:2507.21449.

[6] Timaeus. Guide for Sampling Hyperparameter Selection. 21 April 2026, timaeus.co/research/2026-04-21-sampling-guide.

[7] S. Watanabe. Algebraic Geometry and Statistical Learning Theory. Cambridge University Press, 2009.


# What you would do (about 250 words)

**First, the benchmark.** Four losses with the same Hessian diag(2, 0) at the origin: $u^2$ with an unused v (λ = 1/2, Gaussian calibration); $\tanh^2 u$ with an unused v (λ = 1/2, the regular nonlinear control, Λ − G = 1/(2t)); $u^2 + a v^4$ (λ = 3/4, persistent excess 1/4); $u^2 + b v^6$ (λ = 2/3, excess 1/6). Kept separate: $K_\varepsilon = u^2 + \varepsilon v^2 + v^4$, whose quartic regime $\gamma^2 \ll t \ll \varepsilon^{-2}$ exists only when $\varepsilon\gamma \ll 1$, as a resolution test; and $u^2v^2$, with its 1/2 − 1/log ρ approach, as a slow-convergence test. For each, compute Λ(t, γ) by quadrature on a declared domain, localiser imposed in the stated coordinates, and G on the same domain.

**Second, the diagnostic, fixed before any network result.** Over a declared t-range, fit Λ − G to A + B/t + C/log t; the excess is persistent when A separates from zero at the stated resolution

while the tanh² control returns A = 0. Quantify the minimum detectable curvature ε_min from the K_ε window. Then run SGLD at the same (t, γ), varying step size and chain length separately, to measure $\hat{\lambda} - \lambda$ and the cost of resolving 1/6.

**Third, the network.** Two students with zero truth: the two-unit network at $(a_1, a_2, b_1, b_2) = (1, 2, 0, 0)$, and a four-unit student made of a fixed nonzero two-unit teacher plus a zero pair, (r, c, k) = (5, 6, 3), λ = 8/3, teacher units free. Evaluate the actual loss in the original coordinates with the localiser imposed there; derive the sextic coefficient (1/700 before KL normalisation) and the induced localisation metric; predict the visibility range $t \asymp \gamma^3/a$ and its cost; then test the prediction with the diagnostic of step two.

**Fourth, deferred until the first three work:** one training trajectory with the checkpoint-generated divergence $D(p_{w_t} \| p_w)$ as the loss, stated as such.

# What you expect to see, and what would change your mind (about 150 words)

***Prediction.*** On u² and tanh² the diagnostic returns A = 0, with B = 1/2 for tanh²; on the quartic and sextic controls it returns A = 1/4 and 1/6 once $t \gg \gamma^k/a$; K_ε returns A = 1/4 inside its window and 0 outside; u²v² approaches 1/2 slowly enough that a short t-range misreads it. In the two-unit network the predicted visibility range is nonempty at laptop cost and SGLD returns A = 1/6 within the stated resolution; the four-unit student does the same at higher cost.

***What would make me drop it.*** If the predicted visibility range is empty within the budget, or the diagnostic cannot separate the sextic control from tanh² over any affordable t-range, then higher-order geometry at this order is not measurable with this workflow in this family. The report then names which constraint blocks detection, locality, the finite-temperature correction or the sampler, and how it scales with k and the coefficient. That is a result, not a failure, and it licenses no claim about published LLC curves in other models.

# Scope (about 100 words)

Two weeks give the benchmark, the diagnostic with its false-positive check, and the two-unit network prediction with one test of it. That is the decision point. Three months, only if the two-week criterion is met: the four-unit student, the visibility-and-cost result written as a short paper, and a second geometry (a coinciding pair, k = 2) for the scaling in k. Needs a laptop, the quadrature code from the first proposal's pilot, and devinterp. Drop first the four-unit student, then K_ε. Postponed as separate problems: trajectories, the hidden-noise limit of the layer-composition bound, and complex thresholds of fibre ideals.

# Anything else (optional, about 50 words)

For a mentor: whether Cai's continuous-time targets [4] can be reused directly for u²v²; whether a cutoff domain or a joint (t, γ) limit is the better locality control, given that the Gaussian reference must use the same choice; and whether to merge this with the first proposal above,

which studies the same finite-temperature mechanism from the proximity-to-a-saddle side.