OCTA Research – 365 Days of Learning, Theorizing, and Building AGI

Singular Learning Theory for Transformers: KL Geometry, RLCT, and Attention-Induced Degeneracies

Week 1 · Day 3 · January 3, 2026

OCTA Research Internal Theory Program

Version 1.3 – Living Document Series (Expanded)

OCTA Research 365 Program Note. Day 1 separated macro-level capability "singularity" dynamics from micro-level singularities in Transformer architectures. Day 2 analyzed attention as a stratified, piecewise-smooth map with tie manifolds, argmax patterns, block cells, and parameter varieties.

Day 3 builds the statistical bridge: we apply *Singular Learning Theory* (SLT, due to Watanabe) to Transformer models and, in particular, to attention-induced degeneracies. We identify how tie manifolds and parameter varieties generate singularities in the KL landscape, how the *Real Log Canonical Threshold* (RLCT) replaces parameter count as the effective dimension, and how this shapes scaling laws and generalization.

Version 1.3 expands earlier drafts with: (i) a precise statistical setup for language-model-style Transformers; (ii) formal definitions of RLCT via zeta functions; (iii) toy singular vs. regular examples; (iv) explicit attention-driven degeneracy constructions; (v) an OCTA-flavored Transformer toy model where KL, Fisher, and degeneracy can be computed analytically; (vi) a qualitative sketch of local algebraic structure and resolution-of-singularities intuition; (vii) OCTA design rules for singular architectures; (viii) and additional TikZ visualizations of KL valleys, Fisher spectra, and parameter tie varieties.

Abstract

Most classical statistical learning assumes *regular models*, where Fisher information is nonsingular and maximum-likelihood estimators are asymptotically normal. Deep networks, including Transformers, violate these assumptions: they are *singular models* with parameter non-identifiability, degenerate Fisher spectra, and KL minimizers living on manifolds or varieties, not isolated points.

Singular Learning Theory (SLT) provides the correct asymptotic theory. Its key invariant is the *Real Log Canonical Threshold* (RLCT), which replaces the parameter count in describing generalization error and marginal likelihood.

This Day 3 document:

- formalizes a statistical model for Transformer language models and sequence predictors;
- reviews regular vs. singular models and where Transformers sit;
- introduces Watanabe's zeta function and RLCT;

- shows how attention tie manifolds and parameter varieties from Day 2 generate KL singularities;
- links Fisher spectrum structure to attention stratification and block cells;
- analyzes an explicit OCTA-style Transformer toy model exhibiting attention-induced degeneracy and singular KL geometry;
- explains how RLCT controls generalization, Bayesian evidence, and scaling behavior;
- and defines OCTA-flavored empirical protocols for probing RLCT and singular geometry in real models.

The OCTA perspective: **singularity is not a pathology but the natural geometry of overparameterized intelligence.** We intend to measure, design, and ultimately exploit this geometry rather than avoid it.

Contents

List of Figures

1 Statistical Model for Transformer Sequence Predictors

We first pin down the statistical object to which singular learning theory will be applied.

1.1 Transformer as a conditional density model

Let $\mathcal{X}$ be a space of input sequences (e.g. token sequences) and $\mathcal{Y}$ be outputs (next-token, sequence, label, etc.). A Transformer with parameters θ defines conditional probabilities

$$p_\theta(y \mid x), \quad (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

We assume:

- Data are i.i.d. from an unknown joint distribution

$$q(x, y) = q(x)\, q(y \mid x).$$

- The training objective is minimizing expected negative log-likelihood:

$$L(\theta) = \mathbb{E}_{(x,y)\sim q}[-\log p_\theta(y \mid x)].$$

The KL divergence between the true conditional $q(y \mid x)$ and model $p_\theta(y \mid x)$ is:

$$K(\theta) := D_{\mathrm{KL}}(q \| p_\theta) = \mathbb{E}_{x\sim q(x)} \left[D_{\mathrm{KL}} \left(q(\cdot \mid x) \, \| \, p_\theta(\cdot \mid x) \right) \right]. \tag{1}$$

Up to an additive constant depending only on q, minimizing $K(\theta)$ is equivalent to minimizing $L(\theta)$.

1.2 KL minimizer set

Define the set of KL minimizers:

$$\Theta^\star := \{\theta \in \Theta : K(\theta) = \inf_{\phi\in\Theta} K(\phi)\}. \tag{2}$$

In a regular model:

- $\Theta^\star$ is a *single point*;
- the Fisher information at that point is nonsingular.

For Transformers and other deep networks:

- $\Theta^\star$ is generically a manifold or stratified variety (non-identifiable parameters, symmetries, etc.);
- the Fisher information is *degenerate* (has zero eigenvalues) at every point of $\Theta^\star$.

Understanding the geometry of $\Theta^\star$ and the surrounding KL landscape is the central task of SLT in this context.

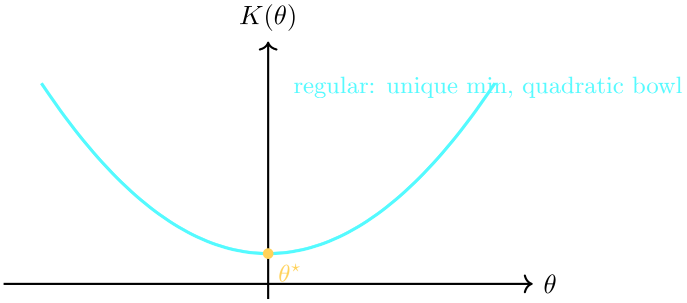

Figure 1: Regular KL geometry: unique minimizer $\theta^\star$ and locally quadratic bowl. Fisher information is nonsingular and parameterization is locally identifiable.

2 Regular vs Singular Models: Geometric Contrast

2.1 Regular case: quadratic KL bowl

In the classical regular case:

- there is a unique $\theta^\star$ minimizing $K(\theta)$,
- the Hessian of K at $\theta^\star$ is positive definite.

Locally:

$$K(\theta) - K(\theta^\star) \approx \frac{1}{2}(\theta - \theta^\star)^\top H(\theta - \theta^\star), \tag{3}$$

where H is the Fisher information matrix.

2.2 Singular case: valleys and cusps

In singular models, there are directions along which $K(\theta)$ is flat or has higher-order vanishing. Toy example in $\mathbb{R}^2$:

$$K(\theta_1, \theta_2) - K(0, 0) \approx \theta_1^4 + \theta_2^2. \tag{4}$$

- Along θ_1-axis: curvature vanishes; the second derivative is zero at the origin.
- The Fisher information degenerates: some eigenvalues are zero.

A more realistic 2D schematic of a singular valley in (θ_1, θ_2) is:
Transformers exhibit exactly this sort of behavior, due to:

- symmetries (neurons/heads permutations, scaling symmetries),
- attention tie manifolds and parameter varieties (Day 2),
- weight matrices whose rank structure collapses.

3 Watanabe's Zeta Function and the RLCT

Singular Learning Theory analyzes the asymptotics of integrals like

$$Z_n = \int_\Theta p_\theta(D_n)\, \pi(\theta)\, d\theta,$$

the marginal likelihood (evidence), where D_n is a dataset of size n and $\pi(\theta)$ is a prior.

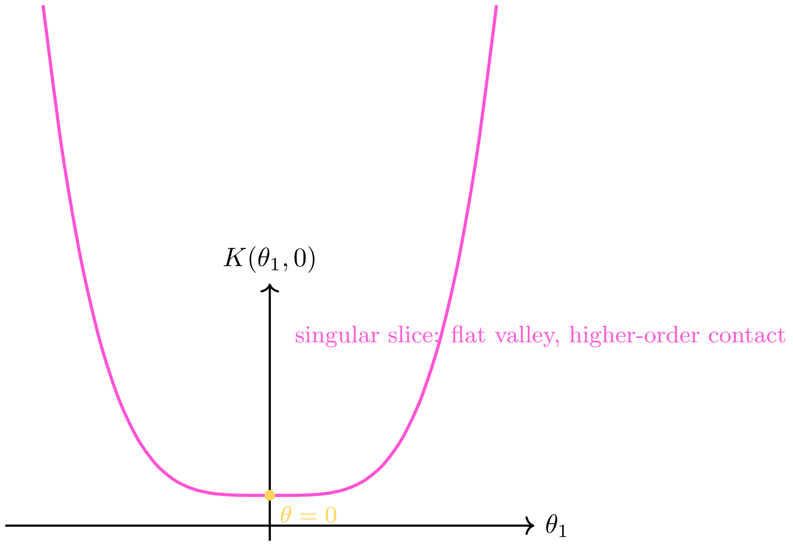

Figure 2: Singular KL geometry (1D slice): higher-order vanishing near the minimum, producing a "flat" valley and degenerate Fisher curvature.

3.1 Kullback–Leibler function and its zeros

Define

$$K(\theta) := D_{\mathrm{KL}}(q \| p_\theta) = \mathbb{E}_q \left[\log \frac{q}{p_\theta} \right].$$

Let

$$K_{\min} := \inf_{\theta \in \Theta} K(\theta), \quad \Theta^\star := \{\theta : K(\theta) = K_{\min}\}.$$

We focus on the behavior of $K(\theta) - K_{\min}$ near $\Theta^\star$.

3.2 Zeta function and RLCT

Following Watanabe, define the *zeta function*:

$$\zeta(z) := \int_\Theta (K(\theta) - K_{\min})^z \, \varphi(\theta) \, d\theta, \tag{5}$$

for a smooth, compactly-supported cutoff $\varphi(\theta)$ around $\Theta^\star$ and complex z.

Definition 3.1 (Real Log Canonical Threshold (RLCT))**.** The *RLCT* λ is the smallest positive real number such that $\zeta(z)$ has a pole at $z = -\lambda$. The order of that pole is called the *multiplicity* m.

Intuitively, λ measures the strength of concentration of the integral near $\Theta^\star$; it plays the role that $d/2$ plays in regular models.

3.3 Asymptotic consequences

For large n,

$$-\log Z_n \;=\; nK_{\min} + \lambda \log n + O(1), \tag{6}$$

and the expected generalization error behaves as

$$\mathbb{E}[G_n] \;=\; \mathbb{E}\left[K(\hat{\theta}_n) - K_{\min} \right] \;\sim\; \frac{\lambda}{n} \quad (n \to \infty), \tag{7}$$

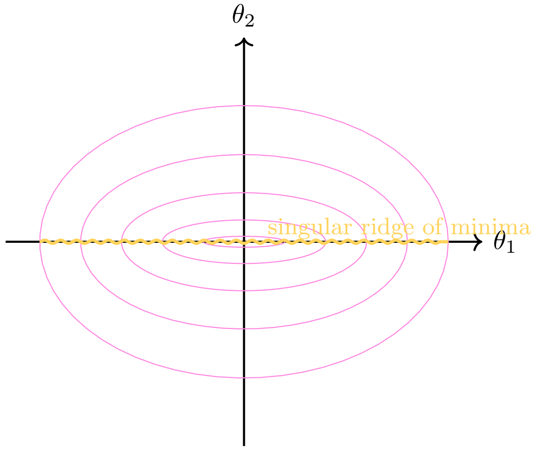

Figure 3: Schematic 2D KL valley with a singular ridge: a continuum of near-minima along θ_1 with stronger curvature in θ_2. This reflects a non-identifiable direction, typical for deep networks.

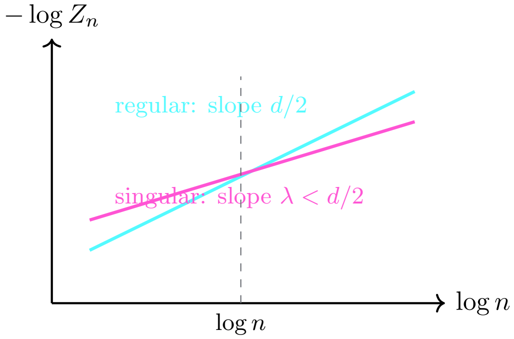

Figure 4: Asymptotic Bayesian free energy $-\log Z_n$ vs. $\log n$. In regular models, the coefficient of $\log n$ is $d/2$; in singular models, it is the RLCT $\lambda < d/2$. The smaller slope reflects effective dimensionality reduction via singular structure.

where $\hat{\theta}_n$ is a maximum likelihood or Bayes estimator.

- In **regular models**, $\lambda = d/2$.
- In **singular models**, $\lambda < d/2$ and depends on the *algebraic structure* of $K(\theta) - K_{\min}$.

From the OCTA viewpoint, λ is a *designable effective dimension* of the model class, shaped by attention geometry and parameter symmetries.

4 Toy Examples: Regular vs Singular RLCT

We recall two canonical examples to fix intuition.

4.1 Regular Gaussian mean

Let $x \in \mathbb{R}$ and q be $\mathcal{N}(\mu_0, 1)$. Model:

$$p_\theta(x) = \mathcal{N}(\mu, 1), \quad \theta = \mu \in \mathbb{R}.$$

Then

$$K(\mu) - K(\mu_0) \propto (\mu - \mu_0)^2,$$

a quadratic bowl. Here:

$$d = 1, \quad \lambda = d/2 = 1/2.$$

4.2 Mixture of Gaussians: classical singular model

Consider a mixture of two Gaussians with unknown weights and means. Different parameter values can represent the same mixture distribution (label-switching symmetry, collapsed components, etc.). The KL minimizer set is not a point, and RLCT $< d/2$. In such models:

- maximum likelihood estimators are not asymptotically normal;
- generalization scales with λ/n where λ is often *fractional*.

Transformers sit firmly in this latter category: many parameters, heavy symmetry, and functional degeneracy.

5 Attention-Induced KL Singularities

Day 2 identified:

- tie manifolds $\mathcal{T}$ in (Q, K)-space;
- parameter tie varieties $\mathcal{V}$ in (W_Q, W_K)-space;
- joint singular set $\mathcal{S}$ in (X, θ)-space.

We now show how these structures produce singularities in the KL geometry.

5.1 Uniform attention degeneracy

Consider a single attention head in a toy classification model. Suppose the value vectors V_j are fixed and the classification layer depends only on the average of Y across tokens.

If $W_Q = 0$ (or $W_K = 0$), then:

$$Q = 0, \quad S_{ij} = 0, \quad A_{ij} = \frac{1}{n},$$

for all (i, j). Any infinitesimal perturbation δW_Q along directions that keep $Q = 0$ on the training set produces *no change* in A, hence no change in $p_\theta(y \mid x)$. Such directions are tangent to the KL minimizer variety.

Locally:

- $K(\theta)$ is constant along these directions zero Fisher curvature;
- the minimum set $\Theta^\star$ contains a manifold parameterized by these symmetries reduced RLCT.

5.2 Key-tying and non-identifiability

Suppose two keys k_j, k_k are always identical for the data distribution (e.g. due to parameter tying or weight initialization). If we also tie the corresponding value vectors $V_j = V_k$, then permuting parameters associated with j and k yields the same function. This generates a finite group symmetry:

$$(\dots, (k_j, V_j), (k_k, V_k), \dots) \mapsto (\dots, (k_k, V_k), (k_j, V_j), \dots),$$

and the quotient in parameter space is nontrivial.

In the space of raw parameters, the manifold of equivalent parameterizations is larger, and KL minima form orbits under such permutations.

5.3 Attention pattern plateaus

From Day 2, within a pattern region $\mathcal{R}_\sigma^\circ$, small parameter changes that keep (Q, K) inside $\mathcal{R}_\sigma^\circ$ (and away from tie manifolds) often change outputs only linearly or even subdominantly. If training has driven the model to a regime where performance is already near optimal for many such regions, the KL surface can be *very flat* in directions that:

- preserve the dominant argmax pattern σ on high-probability data,
- only modulate second-order details of attention weights and values.

Empirically, this translates into:

- long plateaus in the loss landscape,
- many nearly-equivalent minima connected by low-loss paths,
- heavy degeneracy in Hessian and Fisher spectra.

These are signatures of a low RLCT relative to parameter count.

6 Fisher Geometry and Attention Strata

Define the Fisher information matrix:

$$F(\theta) = \mathbb{E}_{(x,y)\sim q}\left[\nabla_\theta \log p_\theta(y \mid x)\, \nabla_\theta \log p_\theta(y \mid x)^\top\right]. \tag{8}$$

6.1 Factorization through attention Jacobians

The gradient factors through the attention Jacobian:

$$\nabla_\theta \log p_\theta = \frac{\partial \log p_\theta}{\partial Z} \frac{\partial Z}{\partial Y} \frac{\partial Y}{\partial A} \frac{\partial A}{\partial S} \frac{\partial S}{\partial (Q, K)} \frac{\partial (Q, K)}{\partial \theta}. \tag{9}$$

Day 2 showed:

- $\partial A / \partial S$ is well-behaved inside pattern regions but can be large/ill-conditioned near tie manifolds;
- inside a fixed strata cell $\mathcal{C}_{\sigma,a}$ (attention pattern σ, MLP activation pattern a), the entire block is smooth and often close to affine.

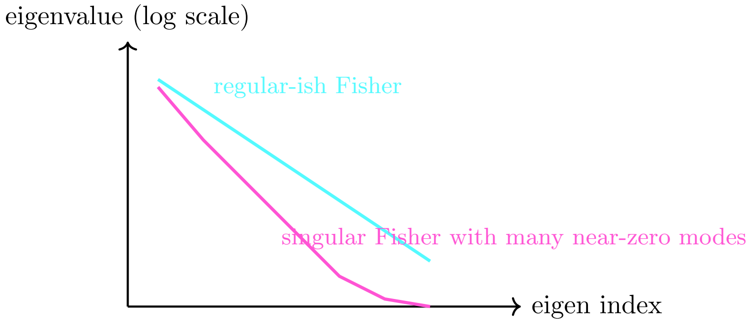

Figure 5: Schematic Fisher spectra. Singular models exhibit many tiny eigenvalues corresponding to directions tangent to KL minimizer manifolds and to near-invariance of attention patterns across strata cells.

6.2 Alignment with strata geometry

Heuristically:

- Eigen-directions with *small eigenvalues* of $F(\theta)$ often correspond to parameter directions that:
  - preserve attention patterns on most data (stay within dominant strata),
  - or move parameters along symmetry or tie varieties.
- Eigen-directions with relatively *large eigenvalues* correspond to directions that:
  - change which strata are visited for typical inputs (cross tie manifolds),
  - or significantly change logits/outputs within each cell.

From an OCTA perspective, Fisher spectra and strata visitation statistics should be co-analyzed: patterns of small vs. large eigenvalues are geometric fingerprints of how the model uses its attention micro-geometry.

7 RLCT as Effective Dimension and Scaling Control

For large n, SLT predicts:

$$\mathbb{E}[G_n] \sim \frac{\lambda}{n}, \tag{10}$$

$$-\log Z_n \sim nK_{\min} + \lambda \log n + O(1). \tag{11}$$

Here λ is typically *fractional* and determined by the local algebraic structure of $K(\theta) - K_{\min}$.

7.1 Comparison to parameter counting and scaling laws

In regular models:

- $2\lambda = d$ is the parameter count;
- learning curves obey $G_n \sim d/(2n)$.

In singular models like Transformers:

- 2λ is often *much smaller* than d;
- this is consistent with empirical scaling laws where effective model capacity seems to grow sublinearly in parameter count;
- from an OCTA design view, we should treat 2λ as the *true capacity* of the model class for a given architecture and data regime.

7.2 OCTA design principle: controlling λ

OCTA Principle 7.1 (RLCT as an OCTA design knob). In OCTA-style architectures, we aim to:

- *shape* the RLCT λ through architectural choices (width, depth, attention head structure, weight tying);
- ensure λ tracks an intended notion of intrinsic complexity (e.g. tasks, Perfect Attractor manifold);
- exploit singularity to gain expressivity without overfitting, via large d but controlled λ.

Roughly:

- more symmetry and redundancy ⇒ smaller λ;
- more independently functional degrees of freedom ⇒ larger λ.

The optimal region likely lies in between: large raw dimension d but moderate effective dimension 2λ.

8 OCTA View: Singular Structure as a Feature

From the OCTA perspective, singularity is fundamental rather than accidental.

OCTA Principle 8.1 (Singularity as Structured Flexibility). Singular models provide entire manifolds of nearly equivalent solutions rather than isolated optima. This structured flexibility:

- enables robust adaptation,
- creates room for multi-objective tradeoffs,
- allows attractors in parameter space to encode families of behaviors instead of points.

In terms of attention and block cells (Day 2):

- each block cell $\mathcal{C}_{\sigma,a}$ corresponds to a micro-regime of computation;
- singularity implies that many parameter settings realize almost the same ensemble of block cells on the data distribution;
- RLCT measures how many such micro-regimes can be flexibly realized without overfitting.

For OCTA, the long-term goal is to:

- map how RLCT changes as we modify attention geometry and head structure;
- engineer architectures where the Perfect Attractor concept appears as a stable low-RLCT pattern of global behavior across tasks.

9 Empirical Protocols for SLT in Transformers

We extend Day 2 protocols with explicit SLT / RLCT probes.

9.1 Protocol D3.1: Generalization curve fitting

Experimental Protocol 9.1 (D3.1: RLCT via generalization curves).

(a) Select a fixed architecture and training procedure.

(b) Train multiple models (or checkpoints) at dataset sizes $n_1 < n_2 < \cdots < n_m$ (e.g. geometric progression).

(c) For each n_k, estimate the generalization gap:

$$G_{n_k} \approx \mathbb{E}_{\text{test}}[-\log p_{\hat{\theta}_{n_k}}(y \mid x)] - H(q),$$

where $H(q)$ is the empirical entropy of labels/targets (may be approximated or treated as a constant offset).

(d) Fit

$$G_{n_k} \approx \frac{\lambda_{\text{eff}}}{n_k} + \frac{b}{n_k^{\alpha}},$$

or, in log form, detect the $1/n$ scaling regime.

(e) Interpret λ_{eff} as an empirical estimate of RLCT (up to constant factors) in the given architecture/data regime.

9.2 Protocol D3.2: Free energy slope via noise tempering

Experimental Protocol 9.2 (D3.2: Evidence slope via temperature/noise).

(a) Consider a Bayesian or approximate Bayesian training scheme (e.g. SWAG, Laplace approximation, SGLD).

(b) Interpret effective dataset size n via inverse noise level or inverse temperature β:

$$\pi_\beta(\theta) \propto p_\theta(D)^\beta \pi(\theta).$$

(c) Estimate $-\log Z(\beta)$ or its surrogate (e.g. negative log-joint at approximate mode plus log-determinant term).

(d) Fit

$$-\log Z(\beta) \approx \beta K_{\min} + \lambda \log \beta + C,$$

to extract an effective λ.

9.3 Protocol D3.3: Fisher–strata coupling

Experimental Protocol 9.3 (D3.3: Fisher eigenvectors and strata visitation).

(a) For a trained model, approximate the Fisher (or Gauss–Newton) matrix $F(\theta)$ at one or several checkpoints (e.g. via Hutchinson trace or low-rank approximations).

(b) Compute leading and trailing eigenpairs (largest and smallest eigenvalues).

(c) For a sample of inputs:

- log which attention patterns σ and block cells $\mathcal{C}_{\sigma,a}$ are visited;
- compute margins to tie manifolds (Day 2 Protocol D2.1).

(d) Correlate:

- low-eigenvalue directions with parameters that mostly preserve strata visitation patterns;
- high-eigenvalue directions with parameters that change which strata are visited or how close typical inputs lie to tie manifolds.

9.4 Protocol D3.4: Parameter sweep across tie varieties

This extends Day 2's parameter sweep (Protocol 2.6) and adds SLT analysis.

Experimental Protocol 9.4 (D3.4: 1D path through parameter singularities).

(a) Choose two trained checkpoints $\theta^{(0)}$ and $\theta^{(1)}$ that differ in attention patterns on a selected validation set.

(b) Define an interpolation path

$$\theta(t) = (1-t)\theta^{(0)} + t\theta^{(1)}, \quad t \in [0,1].$$

(c) For a grid of t values:

- compute validation loss $L(\theta(t))$;
- approximate local curvature (e.g. directional second derivative along the path);
- record pattern distributions $\pi_t(\sigma)$ and Fisher spectra (approximate).

(d) Identify regions along t with:

- sharp pattern switches (crossing tie varieties),
- vanishing curvature (flat plateaus),
- and connect these to local singular structure of $K(\theta)$ (cusps, valleys).

9.5 Protocol D3.5: Architecture-level RLCT comparison

Experimental Protocol 9.5 (D3.5: Effective RLCT across architectures).

(a) Fix a family of architectures with varying:

- number of heads,
- hidden width,
- depth,
- presence/absence of attention masks or weight tying.

(b) For each architecture, perform D3.1 (generalization curve fitting) and estimate λ_{eff}.

(c) Plot $2\lambda_{\text{eff}}$ vs. raw parameter count d.

(d) Identify:

- architectures where $2\lambda_{\text{eff}} \ll d$ (high redundancy, strong singularity),
- architectures where $2\lambda_{\text{eff}}$ grows closer to d (less symmetry).

(e) Use these observations as OCTA feedback for architecture design: choose regions where RLCT aligns with desired complexity and robustness.

10 Explicit OCTA Toy Model with Attention

To make SLT concrete in a Transformer setting, we define a minimal OCTA-style toy model where we can see KL, Fisher, and degeneracy analytically.

10.1 Setup: two-token, one-head, binary classifier

Consider sequences of length $n = 2$ with scalar embeddings ($d_h = 1$) and one attention head. Inputs:

$$x = (x_1, x_2) \in \mathbb{R}^2.$$

We define:

$$q_i = w_Q x_i, \quad k_j = w_K x_j, \quad v_j = w_V x_j,$$

with scalar parameters $w_Q, w_K, w_V \in \mathbb{R}$.

Scores:

$$S_{ij} = q_i k_j = w_Q w_K x_i x_j.$$

For simplicity, we use attention from the first token only and pool its output:

$$A_{1j}^{(\tau)} = \frac{\exp(S_{1j}/\tau)}{\exp(S_{11}/\tau) + \exp(S_{12}/\tau)}, \tag{12}$$

$$Y = \sum_{j=1}^{2} A_{1j}^{(\tau)} v_j. \tag{13}$$

Then a scalar logit is

$$z = w_o Y + b_o,$$

and the model outputs a Bernoulli probability

$$p_\theta(y = 1 \mid x) = \sigma(z), \quad \sigma(z) = \frac{1}{1 + e^{-z}},$$

with $\theta = (w_Q, w_K, w_V, w_o, b_o)$.

10.2 Data distribution

Define a simple data distribution with two sequences:

$$x^{(+)} = (1, 0), \quad y^{(+)} = 1, \tag{14}$$

$$x^{(-)} = (0, 1), \quad y^{(-)} = 0, \tag{15}$$

and

$$q(x^{(+)}, y^{(+)}) = q(x^{(-)}, y^{(-)}) = \frac{1}{2}.$$

Intuitively:

- the first token should be associated with label 1;
- the second token with label 0.

10.3 Model outputs on the two examples

On $x^{(+)} = (1, 0)$:

$$q_1 = w_Q,\ k_1 = w_K,\ k_2 = 0,$$

so

$$S_{11} = w_Q w_K, \tag{16}$$

$$S_{12} = 0. \tag{17}$$

Values:

$$v_1 = w_V, \quad v_2 = 0.$$

Attention weights:

$$A_{11}^{(\tau)} = \frac{\exp(w_Q w_K/\tau)}{\exp(w_Q w_K/\tau) + 1}, \quad A_{12}^{(\tau)} = 1 - A_{11}^{(\tau)}.$$

Output:

$$Y^{(+)} = A_{11}^{(\tau)} w_V.$$

Logit and probability:

$$z^{(+)} = w_o A_{11}^{(\tau)} w_V + b_o, \quad p_\theta^{(+)} = \sigma(z^{(+)}).$$

On $x^{(-)} = (0, 1)$:

$$q_1 = 0,\ k_1 = 0,\ k_2 = w_K,$$

so

$$S_{11} = 0, \quad S_{12} = 0,$$

and attention is uniform:

$$A_{11}^{(\tau)} = A_{12}^{(\tau)} = \frac{1}{2}.$$

Values:

$$v_1 = 0, \quad v_2 = w_V,$$

so

$$Y^{(-)} = \tfrac{1}{2} w_V, \quad z^{(-)} = w_o \tfrac{1}{2} w_V + b_o, \quad p_\theta^{(-)} = \sigma(z^{(-)}).$$

10.4 KL and degeneracy

The KL divergence (cross-entropy up to additive constant) is

$$K(\theta) = \frac{1}{2}\left[-\log p_\theta^{(+)} - \log\Big(1 - p_\theta^{(-)}\Big)\right] + \text{const.} \tag{18}$$

Observe:

- The term $w_Q w_K$ appears in $x^{(+)}$ only via $A_{11}^{(\tau)}$.
- On $x^{(-)}$, attention is independent of w_Q, w_K (pure degeneracy).

In particular:

- If $w_V = 0$ and b_o is tuned so that $p_\theta^{(+)} = p_\theta^{(-)} = \frac{1}{2}$, then *any* (w_Q, w_K) yields the same predictions on both examples.
- The set $\{(w_Q, w_K, w_V = 0, w_o, b_o) : p_\theta^{(+)} = p_\theta^{(-)} = \frac{1}{2}\}$ is a 3D manifold of KL minimizers for this toy dataset.

Thus in the subspace where $w_V = 0$ and (w_o, b_o) are chosen appropriately, the KL landscape is exactly flat in (w_Q, w_K). This is a simple, explicit manifestation of singularity induced by the fact that in this regime the classifier ignores attention completely.

10.5 Attention tie variety

Even away from $w_V = 0$, degeneracies appear when $w_Q w_K \approx 0$ on $x^{(+)}$. Then $S_{11} \approx 0$, $S_{12} = 0$, so attention on $x^{(+)}$ is nearly uniform, similar to $x^{(-)}$.

If we define the parameter tie set

$$\mathcal{V} = \{(w_Q, w_K) : w_Q w_K = 0\},$$

then:

- On $\mathcal{V}$, both $x^{(+)}$ and $x^{(-)}$ have uniform attention.
- The model effectively collapses to a simple linear classifier on w_V (plus biases), independent of (w_Q, w_K).
- Parameter directions that move (w_Q, w_K) along $\mathcal{V}$ leave $K(\theta)$ unchanged.

In this toy model, $\mathcal{V}$ is the union of axes:

$$\{w_Q = 0\} \cup \{w_K = 0\},$$

a simple algebraic variety; in realistic models, such sets are higher-dimensional and more complex but conceptually similar.

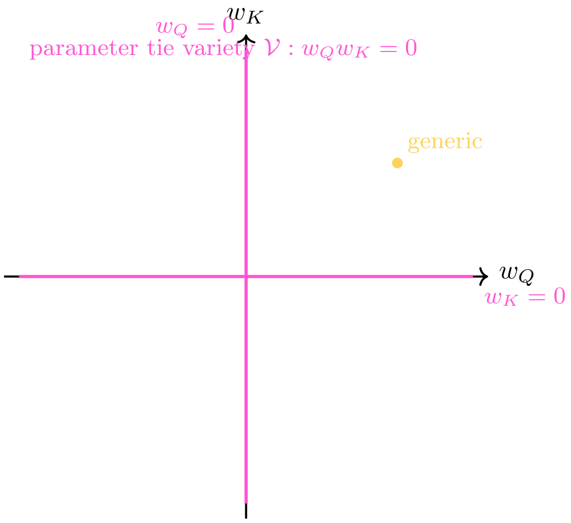

Figure 6: Parameter tie variety in the OCTA toy model: on $\{w_Q w_K = 0\}$, attention becomes uniform on the positive example, collapsing the effective model and creating a manifold of KL-minimizing parameters.

10.6 Fisher degeneracy in the toy model

At a parameter point on $\mathcal{V}$ where $w_V = 0$ and (w_o, b_o) yield $p_\theta^{(+)} = p_\theta^{(-)} = \frac{1}{2}$:

- the gradient of log-likelihood with respect to w_Q and w_K vanishes for both examples,
- the second derivatives in these directions vanish as well (no signal from the data),
- so the Fisher matrix has at least two zero eigenvalues (for w_Q and w_K).

In other words, in this toy example, the Fisher degeneracy and KL singularity can be checked by direct differentiation.

This explicit micro-model serves as a concrete OCTA reference: it shows attention, tie varieties, KL singularity, and Fisher degeneracy all in one analytic object.

11 Local Algebraic Structure and Resolution Intuition

Full algebraic geometry is beyond our scope here, but we can sketch the intuition behind how RLCT arises from the local structure of $K(\theta) - K_{\min}$.

11.1 Monomial-like behavior near singular sets

Locally near a point $\theta^\star \in \Theta^\star$, one often has:

$$K(\theta) - K_{\min} \approx \sum_{j=1}^{r} c_j \prod_{k=1}^{p} |\xi_k(\theta)|^{2a_{jk}}, \tag{19}$$

where:

- $\xi_k(\theta)$ are local analytic coordinates adapted to the singular set,

- a_{jk} are non-negative integers,
- $c_j > 0$ are coefficients.

In the regular case, this reduces to a purely quadratic form (all $a_{jk} = 1$). In the singular case, some exponents are larger ($a_{jk} > 1$) or some coordinates may not appear at all (flat directions).

11.2 Resolution-of-singularities intuition

Algebraic geometry provides formal tools to:

- change variables (blow-ups) to convert $K(\theta) - K_{\min}$ into a simpler monomial form;
- express the integral in the zeta function in terms of integrals of monomials, whose poles and exponents can be read off from the exponents a_{jk}.

Watanabe's theory essentially says:

OCTA Principle 11.1 (RLCT from monomial exponents). After an appropriate analytic change of coordinates and desingularization, the leading pole of the zeta function, i.e. the RLCT λ, can be computed from the exponents of monomials in the local representation of $K(\theta) - K_{\min}$ near the singular set.

In the OCTA toy model of Section 10, near a degenerate manifold like $\{w_V = 0, w_Q w_K = 0\}$, we expect:

- monomial factors like $(w_Q w_K)^2$ to appear in $K(\theta) - K_{\min}$,
- leading to higher-order vanishing in directions normal to the variety,
- and flatness in directions tangent to the variety, all contributing to a reduced RLCT.

11.3 Schematic view in 2D

A schematic 2D picture of a KL function with a monomial-like structure near an axis-aligned singularity is:

The exponents $(4, 2)$ and their combination determine the RLCT in this toy 2D case. In high-dimensional Transformer parameter spaces, similar monomial patterns emerge from rank constraints, tied keys, and attention plateaus.

12 Extended OCTA Design Rules for Singular Architectures

Combining Days 1–3, we can articulate explicit OCTA design rules.

12.1 Design Rule 1: Treat RLCT as a primary hyperparameter

Rather than focusing on raw parameter count d:

- treat 2λ as the *true* complexity of the model class;
- approximate λ empirically using Protocols D3.1 and D3.5;
- adjust architecture (width, depth, head count, tying) to move λ toward a desired target.

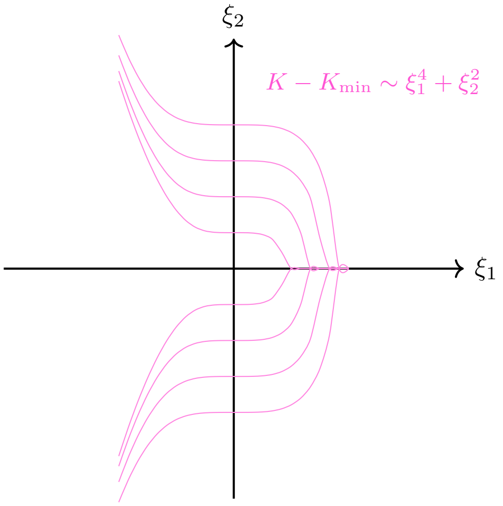

Figure 7: Schematic level sets of a monomial-like KL function $K - K_{\min} \sim \xi_1^4 + \xi_2^2$, illustrating higher-order contact in ξ_1 and quadratic behavior in ξ_2. Such structures appear near attention-induced singular sets after appropriate local coordinate changes.

In particular:

- if the model underfits even at large d, increase λ (reduce redundancy, symmetries);
- if the model overfits despite large n, decrease λ (increase redundancy, add tying, or constrained attention).

12.2 Design Rule 2: Engineer strata geometry

Given the attention stratification of Day 2:

- design head sizes, masks, and key/query projections such that typical data trajectories:
  - remain in a controlled set of strata and block cells;
  - avoid pathological regions where tie manifolds and parameter varieties create fragile behavior;
- tune temperature τ to adjust how sharply transitions between strata happen, trading expressivity against stability.

12.3 Design Rule 3: Monitor Fisher–strata alignment

Using Protocol D3.3:

- track eigenvectors of the Fisher matrix and how they relate to changes in attention patterns;
- design architectures where:
  - key semantic directions (those affecting tasks of interest) correspond to moderately large Fisher eigenvalues;
  - nuisance or purely redundant directions lie in the small-eigenvalue subspace.

This encourages singularity to exist in the "right" directions: those that do not hurt performance or robustness.

12.4 Design Rule 4: Perfect Attractor compatibility

Within the broader OCTA program, we want Perfect Attractors to be:

- *statistically compatible* with model singularities: attractor basins should coincide with regions where RLCT is stable and Fisher spectra are well-behaved;
- *geometrically compatible* with strata structure: attractor trajectories should move across block cells in controlled, interpretable patterns.

This suggests:

- designing training curricula that explicitly move the model through specific strata families while keeping it away from pathological singular loci;
- using RLCT estimates and Fisher diagnostics as "health checks" on whether the attractor structure is aligned with the desired behavior.

13 Roadmap and Relation to Days 1–2

Day 3 completes the Week 1 conceptual triangle:

- **Day 1:** macro-level *singularity* vs. micro-level *singularities*; conceptual separation and toy ODEs;
- **Day 2:** attention as a stratified map on (Q, K) with tie manifolds, block cells, and parameter varieties — the *micro-geometry*;
- **Day 3:** KL geometry, RLCT, Fisher spectra, and singular learning theory — the *statistical geometry.*

The dependencies are:

- attention strata and parameter varieties (Day 2) define the singular set $\mathcal{S}$ in (X, θ);
- this singular set determines the algebraic structure of $K(\theta)$ near $\Theta^\star$;
- the algebraic structure determines RLCT λ and multiplicity m;
- λ and m govern asymptotic generalization and free energy.

Upcoming days can build on this in multiple directions:

- **Day 4:** practical Fisher and Jacobian estimation in large Transformers, plus empirical approximations to RLCT and connection to scaling laws in real models;
- **Day 5:** scaling laws and their fit to SLT predictions, including deviations, phase transitions, and emergent capabilities, particularly around attention and representation stratification;
- **Week 2:** attractors in representation space and how RLCT interacts with Perfect Attractor dynamics in OCTA.

14 Conclusion

Day 3 has:

- formalized the Transformer as a conditional density model $p_\theta(y \mid x)$ with KL divergence $K(\theta)$ as the fundamental loss;
- reviewed regular vs. singular models and illustrated the difference in KL geometry with 1D and 2D schematics;
- introduced Watanabe's zeta function and the Real Log Canonical Threshold (RLCT) as the key invariant governing generalization and free energy;
- clarified how attention tie manifolds, parameter tie varieties, and block cells from Day 2 produce singularities in the KL landscape and degeneracies in Fisher spectra;
- interpreted RLCT as an effective dimension controlling generalization and Bayesian free energy, replacing raw parameter count for singular Transformer models;
- framed singular structure as an OCTA design feature rather than a bug, providing structured flexibility, robustness, and manifold families of nearly equivalent solutions;
- analyzed a concrete OCTA toy model where attention, parameter tie varieties, KL singularity, and Fisher degeneracy can be seen analytically in a minimal setting;
- sketched the local algebraic structure of $K(\theta) - K_{\min}$ and the resolution-of-singularities intuition that underlies RLCT calculations;
- defined a set of empirical protocols (D3.1–D3.5) for probing RLCT-like behavior, Fisher–strata coupling, and architecture-level singular geometry in real models;
- and distilled extended OCTA design rules for singular architectures, tying together RLCT, attention stratification, Fisher geometry, and the Perfect Attractor concept.

In the broader OCTA Research 365-day program, Day 3 is the statistical spine connecting:

- the macro-level notion of capability singularity (Day 1),
- the micro-level attention and block geometry (Day 2),
- and future work on attractors, scaling, safety, and controlled AGI behaviors.

Subsequent days will make these ideas executable at scale: estimating Fisher spectra in large models, measuring effective RLCT from learning curves and Bayesian evidence, and using these signals to steer architectural and training choices toward robust, controllable singular intelligence in the OCTA ecosystem.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry.* AMS and Oxford University Press, 2000.

[2] J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. arXiv:2001.08361, 2020.

[3] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv:1412.6614, 2015.

[4] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

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

[6] S. Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.