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

Training Dynamics on Singular Landscapes: Plateaus, Jumps, and Implicit Bias in Transformers

Week 1 · Day 6 · January 6, 2026

OCTA Research Internal Theory Program

Version 1.4 – Living Document Series (Layout-Refined)

OCTA Research 365 Program Note (Week 1, Day 6).
Days 1–3 established singular geometry as the correct lens for Transformers (macro vs. micro singularities, attention strata, and SLT/RLCT). Day 4 built Fisher-based probes and RLCT proxies. Day 5 lifted these into *singular scaling surfaces* and capability loci over (N, P).

Day 6 asks:

How does SGD move through a singular landscape, and why do we see plateaus, sudden loss drops, and capability jumps?

We argue that:

- plateaus correspond to motion *along* low-curvature singular manifolds;
- sharp drops correspond to *transverse exits* into new strata;
- SGD noise induces an anisotropic diffusion governed by Fisher geometry;
- implicit bias is geometric: flatter, lower-RLCT valleys are preferred;
- capability transitions are *training-time singular events* along the optimization trajectory.

We formalize these statements via projected dynamics, SGD-as-SDE, and OCTA training telemetry.

Contents

List of Figures

1 Introduction: From Static Geometry to Dynamics

Days 3–5 treated singular geometry and scaling as essentially static:

- the parameter space carries a stratified, singular structure;
- RLCT and Fisher spectra encode effective dimension;
- scaling exponents summarize how performance behaves as (N, P) vary.

In practice, models are trained by stochastic gradient methods:

$$\theta_{t+1} = \theta_t - \eta_t \, g(\theta_t; \xi_t), \tag{1}$$

where

- η_t is the learning rate,
- $g(\theta_t; \xi_t)$ is a minibatch gradient from random draw ξ_t.

Day 6 adds the missing dimension: *time*. The trajectory $t \mapsto \theta_t$ is a stochastic curve on a singular manifold. Our objectives are:

1. to relate plateaus and loss jumps to normal vs. tangential components of motion;
2. to connect SGD noise to implicit bias via stationary distributions and curvature;
3. to define OCTA training phases (T0–T3) along time;
4. to propose OCTA telemetry protocols that turn training into a geometry experiment.

2 Projected Gradient Flow on Singular Manifolds

We idealize SGD as a continuous gradient flow:

$$\dot{\theta}_t = -\nabla L(\theta_t). \tag{2}$$

Let $\Theta^\star$ be the set of global minima of L. In singular models, $\Theta^\star$ is not a point but a stratified set:

$$\Theta^\star = \bigcup_s \Theta_s,$$

with each Θ_s a smooth manifold (stratum) of dimension d_s.

2.1 Normal and tangential decomposition

Fix $\theta^\star \in \Theta_s$. Assume locally:

$$T_{\theta^\star}\mathbb{R}^d \cong T_{\theta^\star}\Theta_s \oplus N_{\theta^\star}\Theta_s,$$

with $T_{\theta^\star}\Theta_s$ tangent to the stratum and $N_{\theta^\star}\Theta_s$ a chosen normal space.

Let $P_\parallel$ and $P_\perp$ be the corresponding projection operators. Near $\theta^\star$,

$$\nabla L(\theta) = \nabla L(\theta)^\parallel + \nabla L(\theta)^\perp,$$

with

$$\nabla L(\theta)^\parallel := P_\parallel \nabla L(\theta), \qquad \nabla L(\theta)^\perp := P_\perp \nabla L(\theta).$$

Definition 2.1 (Normal vs. tangential dynamics)**.** The dynamics of (2) decomposes as

$$\dot{\theta}_t^\parallel = -\nabla L(\theta_t)^\parallel, \qquad \dot{\theta}_t^\perp = -\nabla L(\theta_t)^\perp.$$

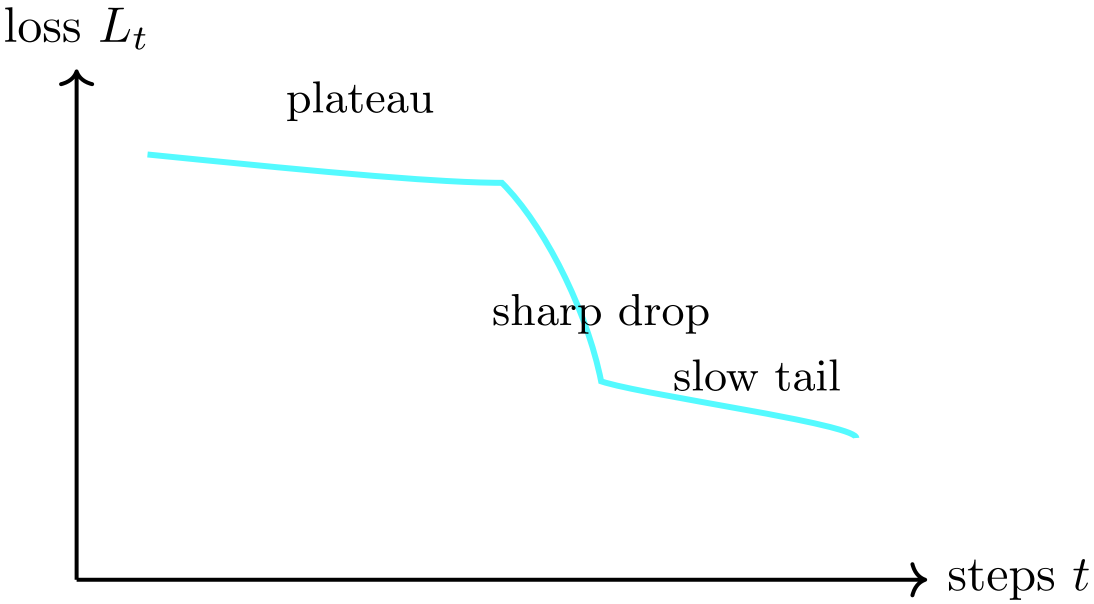

Figure 1: Typical training loss vs. steps: extended plateau, sudden drop, slow tail. Day 6 interprets these qualitatively as dynamics on a singular manifold.

2.2 Linearization near a stratum

Linearize L near $\theta^\star$:

$$L(\theta^\star + \delta\theta) \approx L(\theta^\star) + \frac{1}{2}\delta\theta^\top H^\star \delta\theta, \quad H^\star := \nabla^2 L(\theta^\star).$$

In coordinates adapted to the decomposition:

$$\delta\theta = \begin{pmatrix} \delta\theta_\parallel \\ \delta\theta_\perp \end{pmatrix}, \quad H^\star \approx \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_\perp \end{pmatrix},$$

with $\Lambda_\perp$ positive-semidefinite.

Lemma 2.2 (Leading-order normal contraction)**.** *At leading order near $\theta^\star$,*

$$\dot{\delta\theta}_\perp = -\Lambda_\perp \delta\theta_\perp, \qquad \dot{\delta\theta}_\parallel = 0.$$

If $\Lambda_\perp$ has strictly positive eigenvalues, then $\delta\theta_\perp(t)$ decays exponentially and $\delta\theta_\parallel(t)$ is constant to first order.

Remark 2.3. Higher-order terms in L generate slow, higher-order drift along $T_{\theta^\star}\Theta_s$, but this is suppressed compared to the normal contraction. This separation of time scales underlies the intuitive picture:

- rapid descent toward a valley (normal contraction),
- slow drift along the valley (tangential drift).

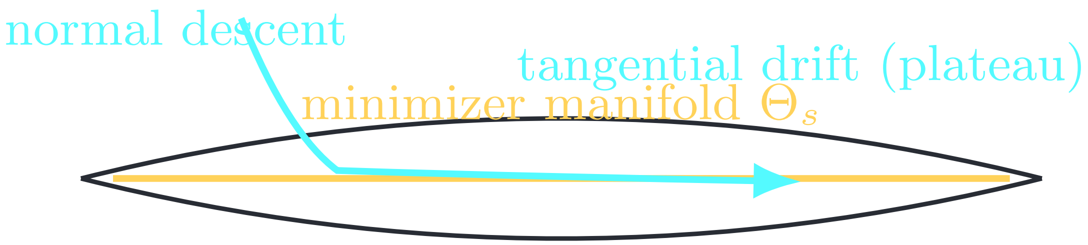

Figure 2: Gradient flow near a singular valley: rapid contraction toward the minimizer manifold, followed by drift along it with minimal loss change.

2.3 Singular drift directions

Definition 2.4 (Singular drift direction)**.** A nonzero vector $v \in T_{\theta^\star}\Theta_s$ is a *singular drift direction* if

$$\nabla^2 L(\theta^\star)v = 0.$$

Along such directions, curvature vanishes at second order; higher-order terms control the behavior:

$$L(\theta^\star + \epsilon v) = L(\theta^\star) + O(\epsilon^k), \quad k > 2.$$

OCTA Principle 2.5 (OCTA Dynamics Principle I – Plateaus as tangential drift)*.* In regions where the trajectory has reached a neighborhood of Θ_s and dynamics is dominated by drift along singular drift directions, we see:

$$\|\dot\theta_t^\perp\| \ll \|\dot\theta_t^\parallel\|, \qquad |\dot L(\theta_t)| \approx 0,$$

which manifests as training plateaus: parameters move, loss barely changes.

3 SGD as Stochastic Differential Equation

Now we reintroduce noise. Write the minibatch gradient as:

$$g(\theta_t; \xi_t) = \nabla L(\theta_t) + \zeta_t,$$

with $\mathbb{E}[\zeta_t \mid \theta_t] = 0$ and covariance

$$\Sigma(\theta_t) := \mathbb{E}[\zeta_t\zeta_t^\top \mid \theta_t].$$

Then update (1) becomes

$$\theta_{t+1} = \theta_t - \eta_t \nabla L(\theta_t) - \eta_t\zeta_t.$$

In the small-step limit, SGD can be approximated by the SDE:

$$d\theta_t = -\nabla L(\theta_t)\,dt + \sqrt{2D(\theta_t)}\,dW_t, \tag{3}$$

where $D(\theta_t)$ is a diffusion tensor depending on $\Sigma(\theta_t)$ and W_t is a Wiener process.

3.1 Normal and tangential SDE components

Decompose the SDE using $P_{\|}$ and $P_{\perp}$:

$$d\theta_t^{\|} = -\nabla L(\theta_t)^{\|} \, dt + \sqrt{2D^{\|}(\theta_t)} \, dW_t^{\|},$$

$$d\theta_t^{\perp} = -\nabla L(\theta_t)^{\perp} \, dt + \sqrt{2D^{\perp}(\theta_t)} \, dW_t^{\perp}.$$

Near Θ_s, $\nabla L(\theta_t)^{\|}$ is small or higher order; thus:

- tangential motion is dominated by noise (diffusion along Θ_s),
- normal motion is dominated by drift (contraction into the valley), with noise occasionally kicking the trajectory away.

OCTA Principle 3.1 (OCTA Dynamics Principle II – Noise-driven exploration). On and near singular manifolds, SGD noise generates an anisotropic diffusion that:

- explores functionally equivalent parameterizations (symmetry directions),
- occasionally discovers directions which, once activated, change the curvature structure and unlock new strata (e.g. new attention heads or circuits).

3.2 Stationary measures and effective temperature

Under further assumptions (constant diffusion, fixed learning rate), the SDE (3) has approximate stationary density:

$$\pi(\theta) \propto \exp\left(-\beta L(\theta)\right),$$

where β is an effective inverse temperature depending on η and noise scale.

On a singular landscape, the stationary measure concentrates on neighborhoods of minimizer manifolds, but with weight modulated by local volume and curvature.

Heuristically:

- high-curvature minima correspond to sharp wells with small volume; they contribute less probability mass;
- flat minima correspond to broader wells with larger volume; they contribute more mass.

OCTA Principle 3.2 (OCTA Dynamics Principle III – Flatness and RLCT bias). In overparameterized Transformers, SGD noise and repeated training implicitly bias toward flatter regions of the minimizer manifolds, which:

- typically exhibit lower effective RLCT (more redundancy, larger singular strata),
- thus produce better generalization (as in SLT) and more favorable singular scaling exponents.

4 Training Plateaus and Training-Time Singular Events

We now formalize plateaus and sharp loss drops as observable phenomena.

4.1 Plateau definition and geometry signature

Let $L_t := L(\theta_t)$ and $\Delta L_t := L_{t+1} - L_t$.

Definition 4.1 (Training plateau (operational))**.** A time window $[t_1, t_2]$ is a *training plateau* if:

(i) the loss changes little:

$$|L_{t_2} - L_{t_1}| \leq \varepsilon_L;$$

(ii) the gradient norm is small on average:

$$\frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \|\nabla L(\theta_t)\| \leq \varepsilon_g;$$

(iii) parameter motion is non-negligible:

$$\|\theta_{t_2} - \theta_{t_1}\| \geq \delta_\theta;$$

for predetermined thresholds $(\varepsilon_L, \varepsilon_g, \delta_\theta)$.

Remark 4.2. Condition (iii) separates plateaus from simply halting training; parameters keep moving due to tangential drift and noise, even if loss is almost constant.

Geometrically, we expect in a plateau:

- Fisher spectra shape stable over time;
- effective dimension $d_{\text{eff}}(t)$ and RLCT proxy $\lambda_{\text{eff}}(t)$ roughly constant;
- head/layer Fisher energies showing slow shifts rather than abrupt changes.

4.2 Training-time singular events: loss and geometry shocks

Definition 4.3 (Training-time singular event)**.** A time $t^\star$ is a *training-time singular event* if:

(i) the loss change is a significant negative outlier:

$$\Delta L_{t^\star} \leq \mu_{\Delta L, t^\star} - k\sigma_{\Delta L, t^\star},$$

with $\mu_{\Delta L,t}, \sigma_{\Delta L,t}$ an online mean/std estimate and k a threshold;

(ii) at the same time, at least one geometry observable experiences a shock, e.g.

$$|\mathrm{Tr}(F_{t^\star}) - \mathrm{Tr}(F_{t^\star - 1})| > \tau_F,$$

or a sharp change in smallest non-zero eigenvalue, RLCT proxy, or headwise Fisher energy.

OCTA Principle 4.4 (OCTA Dynamics Principle IV – Loss drops as transverse exits). Training-time singular events where loss drops sharply and geometry observables jump are interpreted as:

- transverse exits from one stratum Θ_s into another $\Theta_{s'}$,
- accompanied by a change in curvature and effective RLCT,
- often correlated with capability transitions (activation of new behaviors).

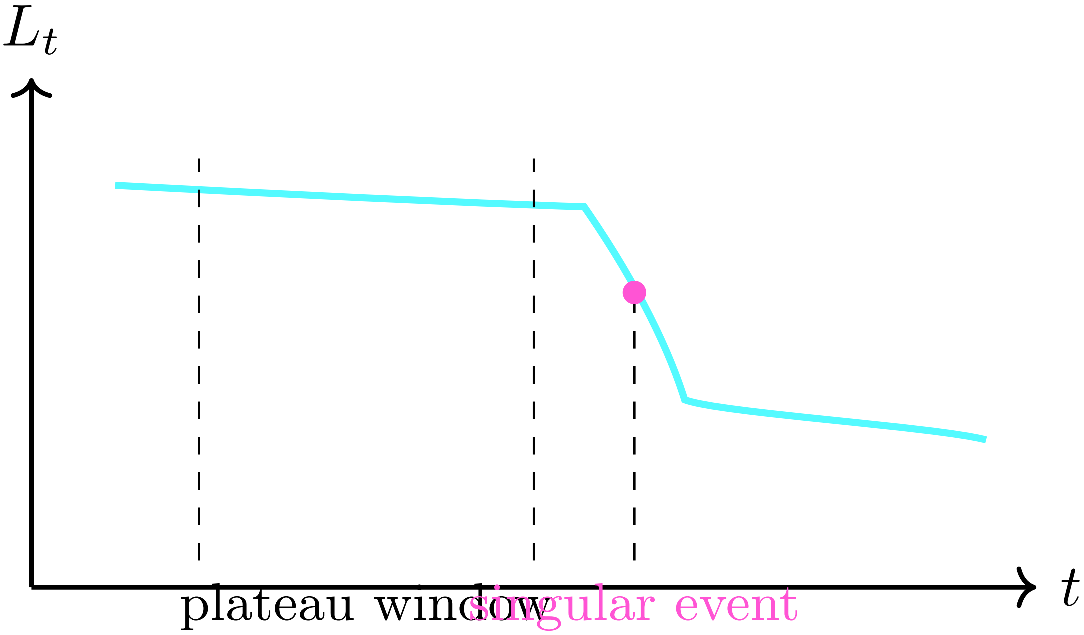

Figure 3: Plateau followed by a training-time singular event (sharp loss drop). OCTA logs geometry metrics at such events to detect phase transitions.

5 Toy Model: Singular Valley with Noise

We now revisit a simple redundant parametrization to illustrate plateaus and jumps qualitatively.

5.1 Redundant parametrization $w = uv$

Consider scalar regression with prediction $\hat{y} = uvx$, squared loss, and data such that the optimal effective weight is $w^\star$. The loss is:

$$L(u, v) = \frac{1}{2}(uv - w^\star)^2.$$

The minimizer set is the curve

$$\Theta^\star = \{(u, v) : uv = w^\star\}.$$

This is a classic singular set: one-dimensional in (u, v)-space.

5.2 Noisy gradient descent dynamics

The gradients are:

$$\partial_u L = (uv - w^\star)v, \qquad \partial_v L = (uv - w^\star)u.$$

Discrete-time noisy gradient descent:

$$u_{t+1} = u_t - \eta(uv - w^\star)v_t + \sqrt{2\eta\sigma^2}\,\xi_t^{(u)},$$

$$v_{t+1} = v_t - \eta(uv - w^\star)u_t + \sqrt{2\eta\sigma^2}\,\xi_t^{(v)},$$

with $\xi_t^{(u)}, \xi_t^{(v)} \sim \mathcal{N}(0, 1)$.

Near $\Theta^\star$, the dynamics splits into:

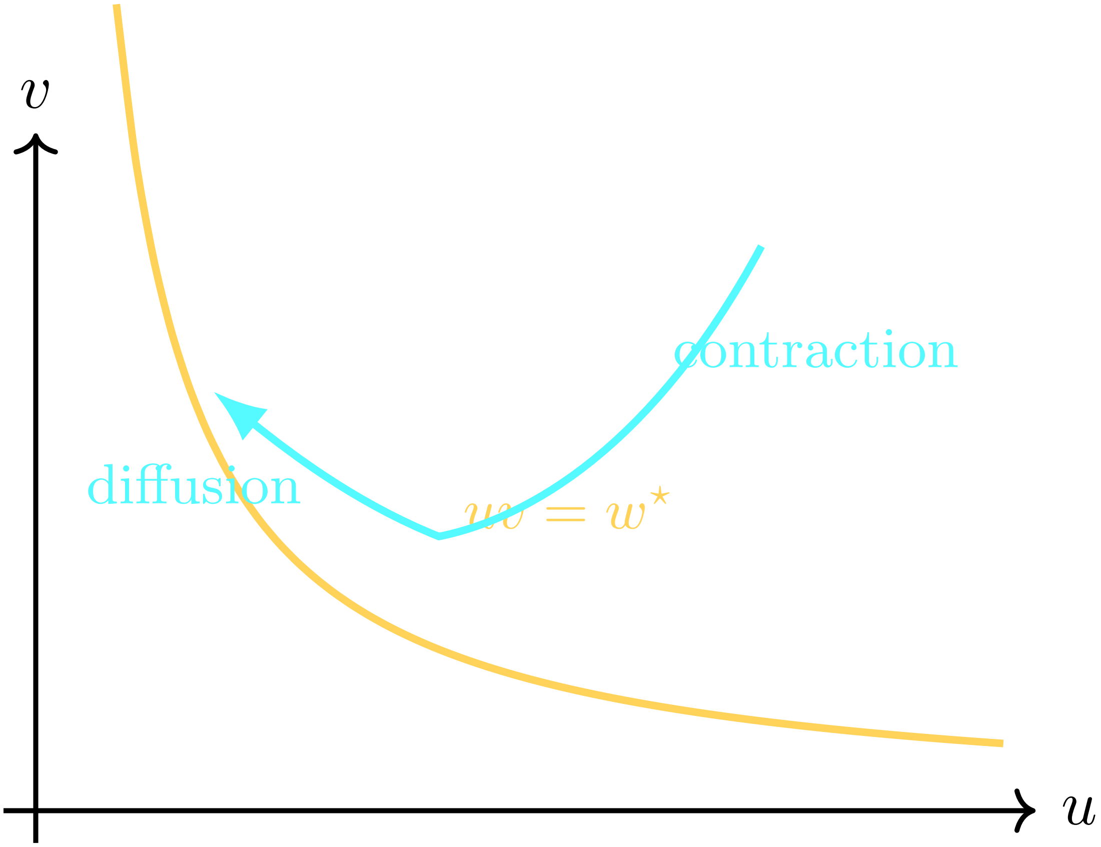

Figure 4: Toy singular valley in (u, v) space: rapid contraction to the curve $uv = w^\star$, then stochastic drift along it due to noise.

- fast contraction toward $uv = w^\star$ (normal to the curve),
- slow stochastic diffusion along $uv = w^\star$ (tangential).

This captures:

- **plateaus:** once near $uv = w^\star$, loss changes little while (u, v) drift;
- **implicit bias:** some (u, v) pairs are favored by noise structure (e.g. smaller norms);
- **event potential:** additional terms or constraints could make some regions of $\Theta^\star$ gateways to new behavior.

In Transformers, analogous redundancies exist:

- multiple heads approximating the same pattern;
- rescaling between layers, projections, and MLPs;
- tied symmetries across blocks.

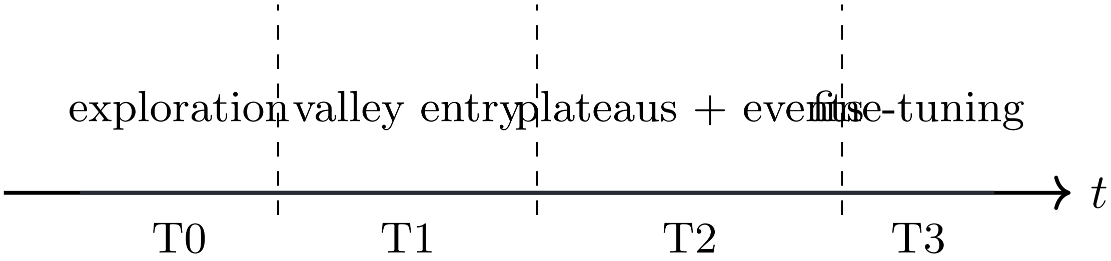

Figure 5: OCTA training phases on the time axis: T0 (exploration), T1 (valley entry), T2 (plateaus + events), T3 (fine-tuning).

6 OCTA Training Phases on the Time Axis

Day 5 defined scaling phases across (N, P) (Phase 0–3). Day 6 defines *training phases* along time t for fixed (N, P).

Definition 6.1 (OCTA training phases T0–T3)**.** For a fixed dataset/model configuration, we define:

- **Phase T0 (Exploration / Chaotic Descent):**
  - high loss, large gradient norm;
  - Fisher spectra broad and unstable;
  - frequent parameter sign changes and head role swaps;
  - rapid changes in attention patterns and token-level behavior.
- **Phase T1 (Valley Entry / Stabilization):**
  - substantial loss decrease;
  - Fisher spectra begin to stabilize in shape (though magnitudes still change);
  - layers and heads start to acquire distinct roles;
  - early plateaus may appear on some metrics.
- **Phase T2 (Plateau + Events):**
  - extended plateaus with tangential drift along singular manifolds;
  - intermittent training-time singular events (loss drops, geometry shocks);
  - associated capability jumps for many tasks.
- **Phase T3 (Fine-Tuning / Saturation):**
  - small incremental gains;
  - geometry and capabilities approach asymptotic profiles;
  - low-frequency minor singular events, if any.

OCTA Principle 6.2 (OCTA Dynamics Principle V – Phase-aware analysis)*.* All training analytics (loss, geometry, capabilities) must be interpreted conditionally on the training phase T0–T3. Aggregating across phases obscures the distinctive dynamics and geometry of each phase.

7 Training Telemetry: Geometry Over Time

We now specify the telemetry OCTA should record during training and how to analyze it.

7.1 Core telemetry signals

At checkpoints $t \in \mathcal{T}$, log:

- **Loss and capabilities:**
  - $L_t^{\text{train}}, L_t^{\text{val}}$,
  - capability metrics $\text{Cap}_T(t)$ for tasks T from Day 5.
- **Fisher geometry:**
  - Fisher trace $\text{Tr}(F_t)$,
  - top-k eigenvalues $\lambda_{t,1}, \ldots, \lambda_{t,k}$,
  - spectrum entropy $H_{\text{spec}}(F_t)$,
  - smallest non-zero eigenvalue $\lambda_{t,\min}^{+}$.
- **Head and layer metrics:**
  - headwise Fisher energies $E_{t,h}$,
  - layerwise traces $T_{t,\ell}$,
  - attention-strata occupancy from Day 2.
- **RLCT / dimension proxies:**
  - approximate RLCT $\lambda_{\text{eff}}(t)$ from local learning-curve fits,
  - curvature-based effective dimension $d_{\text{eff}}(t)$.

7.2 Multichannel telemetry visualization

7.3 Protocols D6.1–D6.4

Experimental Protocol 7.1 (D6.1: Training-phase annotation).

(a) Partition training into windows (by wall-clock or steps).

(b) For each window, compute summary statistics:

- mean and variance of loss change ΔL_t,
- mean gradient norm,
- variance of $\text{Tr}(F_t)$ and $H_{\text{spec}}(F_t)$,
- number and magnitude of singular events (see D6.2).

(c) Assign each window a phase label T0–T3 via heuristic criteria (thresholds on these statistics).

(d) Record phase labels and use them as conditioning variables in scaling and geometry analyses.

Experimental Protocol 7.2 (D6.2: Singular event detector).

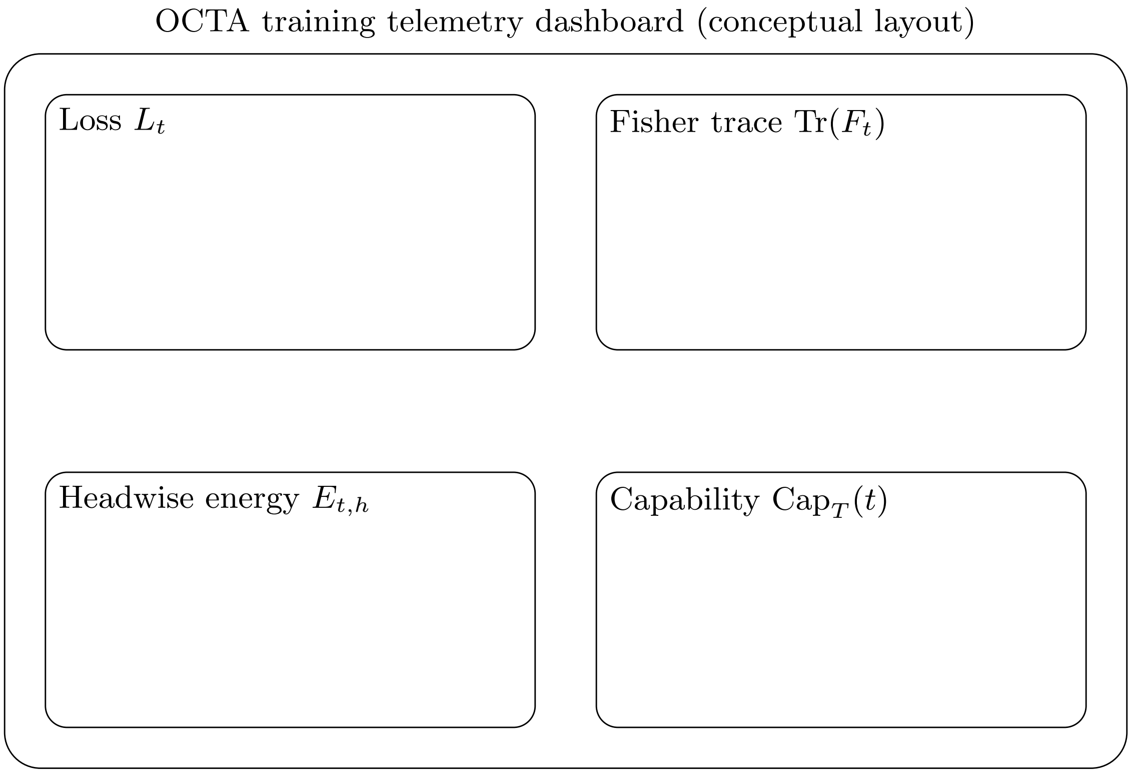
OCTA training telemetry dashboard (conceptual layout)
Loss L_t
Fisher trace $\mathrm{Tr}(F_t)$
Headwise energy $E_{t,h}$
Capability $\mathrm{Cap}_T(t)$

Figure 6: Conceptual OCTA training telemetry dashboard: synchronized views of loss, Fisher trace, headwise energies, and task capabilities over time.

(a) Maintain exponential moving averages $(\mu_{\Delta L,t}, \sigma_{\Delta L,t})$ of ΔL_t.

(b) At each step, flag a *loss event* if

$$\Delta L_t \leq \mu_{\Delta L,t} - k\sigma_{\Delta L,t}$$

for a chosen k (e.g. $k = 3$).

(c) Cross-check geometry observables:

- if any change exceeds its threshold (e.g. Fisher trace jump $> \tau_F$), classify as a *training-time singular event*;
- otherwise, classify as a noise fluctuation.

(d) For singular events, log:

- pre- and post-event loss and capabilities,
- pre- and post-event geometry (Fisher, RLCT proxy, head/layer metrics),
- training phase context (T0–T3).

Experimental Protocol 7.3 (D6.3: Plateau characterization).

(a) Detect plateaus using the definition in Section 4.

(b) For each plateau window:

- estimate average Fisher spectrum and entropy;
- estimate RLCT proxy and effective dimension;
- summarize drift distance in parameter space and principal drift directions;
- record head/layer energy distributions and attention-strata statistics.

(c) Compare plateaus across training runs and architectures to identify:

- typical geometry profiles in T2,
- correlations between plateau geometry and subsequent capabilities,
- architectures with more "useful" plateaus (leading to beneficial events).

Experimental Protocol 7.4 (D6.4: Early-warning geometry monitor).

(a) At each checkpoint, compute:

- $\mathrm{Tr}(F_t)$, $H_{\mathrm{spec}}(F_t)$,
- $\lambda^+_{t,\min}$,
- headwise energies $E_{t,h}$,
- RLCT proxy $\lambda_{\mathrm{eff}}(t)$.

(b) Fit local trends (e.g. moving linear regressions) for these signals.

(c) Trigger alerts when:

- curvature grows sharply (large positive trend in $\mathrm{Tr}(F_t)$ or $\lambda^+_{t,\min}$),
- eigenvalue entropy collapses (Fisher mass concentrates into few directions),
- RLCT proxy exhibits a jump,
- several heads simultaneously increase $E_{t,h}$.

(d) Use alerts to:

- schedule more frequent checkpoints,
- optionally adjust learning rates or regularization,
- perform targeted interpretability analysis around the event.

8 Curvature-Aware Schedules and RLCT Trajectories

Having explicit curvature and RLCT proxies over time enables OCTA to design feedback-based training schedules.

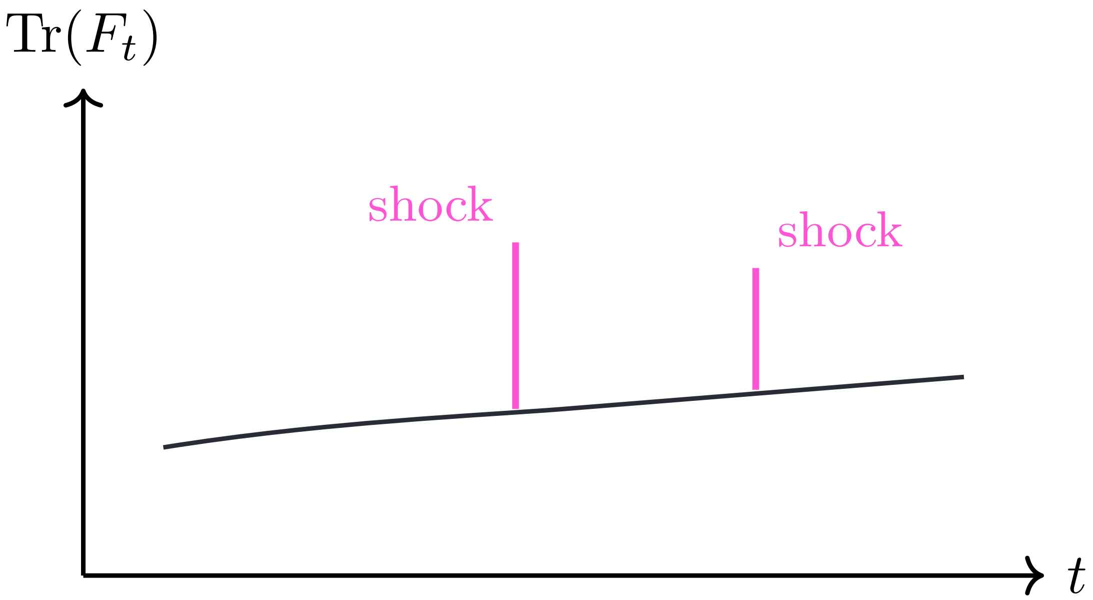

Figure 7: Example Fisher trace over training time with two curvature shocks. D6.4 treats such shocks as early-warning signals for training-time singular events and phase changes.

8.1 Effective RLCT trajectory

Define $\lambda_{\mathrm{eff}}(t)$ as a time-dependent RLCT proxy, estimated via:

- local learning-curve fits on short windows,
- Fisher-based effective dimension and Day 4 methods,
- or combined proxies (e.g. mixing curvature and generalization gap estimates).

We can view $\lambda_{\mathrm{eff}}(t)$ as a coarse measure of how many effective degrees of freedom are being used at time t.

OCTA Principle 8.1 (OCTA Dynamics Principle VII – RLCT trajectory monitoring). The trajectory $t \mapsto \lambda_{\mathrm{eff}}(t)$ provides:

- evidence of when the model has moved into richer functional regimes,
- a way to distinguish meaningful phase transitions from mere noise,
- a control signal for learning rate and regularization schedules.

8.2 Curvature-aware learning rate schedule

Conceptually, one can define a curvature-aware schedule:

$$\eta_{t+1} = \eta_t \cdot f\big(\mathrm{Tr}(F_t), \lambda^{+}_{t,\min}\big),$$

where f decreases η_t when curvature becomes large (to prevent instability) and possibly increases it in stable plateau regions to encourage exploration along flat directions.

Experimental Protocol 8.2 (D6.5: Curvature-aware schedule (conceptual)).

(a) For each checkpoint, compute curvature indicators:

$$C_t := \alpha_1 \mathrm{Tr}(F_t) + \alpha_2 \lambda^+_{t,\min},$$

for chosen weights (α_1, α_2).

(b) Maintain a target curvature range $[C_{\min}, C_{\max}]$.

(c) Update learning rate:

$$\eta_{t+1} = \begin{cases} \eta_t \cdot (1+\epsilon), & C_t < C_{\min} \\ \eta_t, & C_{\min} \le C_t \le C_{\max} \\ \eta_t \cdot (1-\epsilon), & C_t > C_{\max} \end{cases}$$

for small ϵ (e.g. 1% adjustments).

(d) Evaluate whether this schedule:

- reduces harmful curvature spikes,
- preserves beneficial singular events,
- improves stability near capability transitions.

9 Transformer-Specific Interpretation

We now connect these dynamics more concretely to Transformer internals, building on Days 2, 4, and 5.

9.1 Head activation and specialization indices

Let $E_{t,h}$ be headwise Fisher energy for head h at step t. Define:

$$\tilde{E}_{t,h} := \frac{E_{t,h}}{\sum_{h'} E_{t,h'}},$$

and a head specialization entropy:

$$H_{\text{head}}(t) := -\sum_h \tilde{E}_{t,h} \log \tilde{E}_{t,h}.$$

- High $H_{\text{head}}(t)$: energy spread across many heads (less specialization).
- Low $H_{\text{head}}(t)$: energy concentrated in fewer heads (more specialization).

Empirically, we expect:

- in T0/T1: $H_{\text{head}}(t)$ relatively high and noisy;
- in T2: $H_{\text{head}}(t)$ decreases as some heads become dominant and specialized;
- training-time singular events often coincide with sharp local changes in $H_{\text{head}}(t)$.

OCTA Principle 9.1 (OCTA Dynamics Principle VIII – Head specialization). Head specialization dynamics, as measured by $H_{\text{head}}(t)$ and $E_{t,h}$, are key indicators of:

- movement between strata in attention-function space,
- imminent capability transitions (e.g. longer-range reasoning),
- changes in effective RLCT and scaling behavior.

9.2 Layer role shifts

Similarly, we can define layerwise contribution metrics such as:

$$\tilde{T}_{t,\ell} := \frac{T_{t,\ell}}{\sum_{\ell'} T_{t,\ell'}}.$$

Changes in the profile $\{\tilde{T}_{t,\ell}\}_\ell$ over time can signal:

- early layers stabilizing while later layers remain volatile;
- emergence of mid-layer "bottlenecks" associated with reasoning or planning;
- architectural mismatches when certain depth ranges remain underutilized.

9.3 Capabilities as attractors in singular dynamics

Day 5 defined capability loci $\mathcal{C}_T(\tau)$ in (N, P). At fixed (N, P), Day 6 adds time:

$$\mathrm{Cap}_T(t)$$

often shows:

- flat or noisy behavior in T0/T1,
- step-like increases around training-time singular events in T2,
- saturation in T3.

OCTA Principle 9.2 (OCTA Dynamics Principle IX – Capability singular events). For a given capability T, a *capability singular event* is a training-time singular event $t^\star$ such that:

$$\mathrm{Cap}_T(t^\star) - \mathrm{Cap}_T(t^\star - 1)$$

exceeds a threshold and aligns with geometry shocks (Fisher, RLCT, head/layer metrics). These events mark the emergence of qualitatively new behavior modes (e.g. in-context learning onset).

10 Synthesis with Days 1–5

Day 6 completes the Week 1 arc of OCTA Research:

- **Day 1** separated macro Singularity narratives from micro architectural singularities.
- **Day 2** described attention and block cells as stratified geometric objects.
- **Day 3** introduced SLT and RLCT as the correct asymptotic descriptors.
- **Day 4** gave practical Fisher-geometry and RLCT probes.
- **Day 5** lifted these into singular scaling surfaces and capability loci over (N, P).
- **Day 6** treats training as a stochastic flow on this singular geometry, explaining plateaus, jumps, and implicit bias.

Model architecture ⇒ Singular geometry ⇒
(RLCT, Fisher, scaling surfaces;
SGD dynamics, plateaus, singular events)
⇒ Capabilities & Safety Windows.

From an OCTA perspective:

- scaling laws, training dynamics, and capabilities are not separate stories, but different projections of the same singular geometry;
- telemetry and protocols D6.1–D6.5 promote training to a first-class scientific experiment in singular dynamics, not just an engineering process;
- future days will use this as a scaffold for regularization design, interpretability, and safety.

11 Outlook

Day 6 suggests several immediately actionable directions for OCTA Research:

- **Regularization as geometry shaping** (Days 7–8):
  - analyze how weight decay, dropout, and architectural choices reshape $\Theta^\star$;
  - study how they alter RLCT trajectories and Fisher spectra;
  - design "geometry-aware" regularizers that control the frequency and magnitude of training-time singular events.
- **Interpretability and circuit emergence** (Days 9–10):
  - link geometry shocks and head/layer shifts to interpretable circuits;
  - map capability singular events to concrete patterns (e.g. new attention motifs);
  - build tools to visualize how singular dynamics rearrange internal computation.
- **Safety and monitoring**:
  - define safe and unsafe regions in training-time geometry (analogous to Day 5 safety windows on (N, P));
  - use D6.4 early-warning monitors to detect risky geometry transitions;
  - develop playbooks for interventions during or after strong singular events.

Learning is movement through singular geometry over time;
OCTA's job is to instrument, steer, and understand that movement.

References

[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 1998.

[2] S. Mandt, M. D. Hoffman, D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 2017.

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

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