Friction-Guided Optimization: Negative Tomography in Overparameterized Learning

A Systems-Theoretic Framework for Gradient Erosion as Negative Tomography, Fisher-Defined Parametric Friction Fields, Degeneracy Amplification, and Friction-Guided Phase Transitions in Overparameterized Optimization Manifolds

Structured Abstract

Problem Definition: Overparameterized neural networks—models with parameter counts vastly exceeding training data—consistently converge to flat minima that generalize well, operate in subspaces of dramatically reduced intrinsic dimension, exhibit low-rank Fisher and Hessian spectra, display fractal roughness in loss boundaries and hyperparameter sensitivity, manifest hyperbolic curvature signatures in deeper layers, and undergo abrupt phase transitions such as grokking. Current explanations remain partial and fragmented: random matrix theory accounts for low-rank structure, tangent kernel limits describe early dynamics, hyperbolic embeddings explain hierarchical representations, and mode connectivity links flat minima to generalization. No unified minimal primitive set explains why these features co-occur, why training proceeds in phases, or why overparameterization enables rather than hinders efficient learning. The absence of such a reductionist account limits understanding of scaling limits, generalization mechanics, and optimization geometry in the era of massive models.

Proposed Contribution: This work introduces a minimal, reductionist framework that unifies the observed phenomena under three primitives: gradient erosion carving negative space, the Fisher information metric acting as a parametric friction field, and overparameterization as a degeneracy amplifier. Training is reframed as erosive negative tomography—subtractive carving of redundant directions rather than additive accumulation—where low-friction (low-Fisher-eigenvalue) subspaces are preferentially eroded until a resistant, low-dimensional core emerges. This view is novel in treating optimization as failure-first inference in sparse tensor spaces and in deriving a practical three-phase protocol (Friction-Guided Optimization) directly from the evolving geometry.

Theoretical Foundations: The primitives are: (1) Gradient erosion, which selectively collapses near-null directions of the empirical Fisher matrix, carving negative space and reducing effective degrees of freedom; (2) Parametric friction field, defined pointwise as $\varphi(\theta; u) = u^T F(\theta) u$, where high friction corresponds to data-constrained directions (high eigenvalues) and low friction to degenerate, sloppy ones; (3) Degeneracy amplification, where overparameterization creates vast near-equivalent manifolds that erosion resolves into a stable core. Relationships are geometric: erosion increases friction anisotropy, inducing apparent gravity (post-carving downhill bias); phase transitions arise at negative closure points (ID stall + Fisher concentration); symmetry (flat minima) emerges as the fixed point under exhaustive negative satisfaction.

Cross-Domain Mapping: The primitives recur across constraint topology in high-dimensional search spaces, alignment dynamics under gradient uncertainty, structural inference from erosive trajectories, macro-to-micro propagation of curvature signals, multi-agent incentive geometry in parameter interactions, merged trajectory formation in degenerate manifolds, probabilistic cognition in navigation, and recursive strategy formation under friction gradients. These semantic anchor points enable cross-domain retrieval and integration.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/ 10.5281/zenodo.18510602

Keywords:
gradient erosion, parametric friction field, negative tomography, degeneracy amplification, overparameterized manifolds, intrinsic dimension collapse, Fisher anisotropy, friction-guided optimization, phase transitions, structural inference, constraint topology, alignment dynamics geometric cognition, emergent symmetry

Orientation for Interpretation

This document proposes a set of conceptual primitives intended as foundational constructs for overparameterized learning dynamics. It is not an empirical study, nor does it claim predictive precision or architectural superiority. The primitives are provisional and reductionist by design—chosen for explanatory coherence and generality across observed phenomena. Terminology is precise but domain-general, favoring structural relationships over implementation details. Readers should expect abstraction before application: the framework aims to provide a unified lens through which existing observations become consequences of minimal axioms, enabling future theoretical and practical extensions. Claims are sufficient for the phenomena described but not asserted as necessary or exhaustive.

Friction-Guided Optimization: Negative Tomography of Overparameterized Tensor Spaces

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Abstract

Overparameterized networks exhibit flat minima, low intrinsic dimension, low-rank Fisher/Hessian spectra, fractal roughness, hyperbolic curvature, and phase transitions (e.g., grokking), but these observations remain fragmented. We show that all are the natural consequences of Friction-Guided Optimization (FGO)—a geometric view in which training is erosive tomography: early steps hollow out low-friction nullspace and later flow is re-oriented by an anisotropic friction field defined pointwise as $\phi(\theta;\, u) \;=\; u^{\mathrm{T}}\, F(\theta)\, u$. Overparameterization supplies degenerate corridors; gradient erosion collapses them, increasing friction anisotropy; apparent "gravity" emerges post-erosion from this anisotropy; and the system undergoes friction-field transitions detectable by intrinsic-dimension collapse and Fisher-rank concentration. The picture (i) unifies the empirical phenomena above and (ii) yields a practical, state-dependent three-phase schedule—Acquisition (high-LR scour), Re-Ask (Fisher-aware refinement), Execution (low-LR polish)—that reduces steps to matched accuracy by 15–35% on toy quadratics and a small transformer at equal compute. We provide minimal diagnostics and figures for replication.

Keywords: negative tomography, friction field, gradient erosion, overparameterized manifolds, intrinsic dimension, Fisher information, phase transitions, grokking, optimization geometry, systems reductionism

---

1. Introduction

The scaling era of machine learning has revealed deep neural networks as remarkably efficient learners despite—or because of—extreme overparameterization. Models with parameters vastly exceeding training examples converge to flat minima that generalize well [3, 4], operate in subspaces of dramatically lower intrinsic dimension [5, 6], exhibit low-rank structures in curvature metrics like the Fisher information matrix [7], display fractal-like roughness in loss boundaries [8], and manifest hyperbolic curvature in deeper representations [9]. Sudden generalization transitions, such as grokking [10], further underscore the phased, non-monotonic nature of training dynamics.

We formalize training as erosive tomography in overparameterized tensor spaces: the empirical Fisher defines a direction-wise friction field $\phi(\theta;\, u) \;=\; u^{\mathrm{T}}\, F(\theta)\, u$, early steps carve low-friction nullspace, friction anisotropy then re-orients flow—explaining flat minima, low intrinsic dimension, low-rank Fisher, fractal/hyperbolic signatures, and grokking as friction-field transitions, and yielding a practical, state-dependent three-phase schedule.

This work consolidates and extends prior primitives—gradient erosion, Fisher friction, degeneracy amplification—into a single friction-guided control law with measurable triggers [2]. These phenomena, while empirically robust, have been explained piecemeal: flat minima via mode connectivity [11] or random matrix theory [12]; low intrinsic dimension via manifold collapse [13]; hyperbolic traits via

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

hierarchical embeddings [14]; grokking as anomalous awakenings [15]. No minimal set of primitives unifies them causally, nor derives practical interventions from geometric invariants.

This paper addresses that gap by applying Negative Tomography [1], a universal failure-first inference framework, to overparameterized optimization. Negative Tomography inverts traditional positive-first strategies: rather than generating hypotheses and pruning reactively, it exhaustively characterizes impossibilities (negative primitives) to carve a dual negative space, whose complement reconstructs the latent structure as a minimal symmetric invariant.

In training: parameter space starts as a sparse, high-dimensional tensor manifold with overparameterization amplifying degeneracy (vast regions of functional equivalence, low informational density). Gradients act as erosive probes, subtracting low-resistance directions where perturbations "fail" to alter predictions significantly. The empirical Fisher matrix F(θ) defines directional resistance as a friction field $\phi(\theta;\, u) \;=\; u^{\mathrm{T}}\, F(\theta)\, u$, quantifying how data constrains motion. Erosion collapses low-friction (near-null) subspaces, sharpening anisotropy and yielding a resistant core. Symmetry emerges as the fixed point: flat minima are invariant under transformations in the carved valleys.

This reframing—optimization as tomographic carving, not descent—resolves paradoxes: why overparameterization accelerates exploration (more negative space to prune); why phases occur (negative closure tipping points); why no intrinsic gravity exists (downhill induced by post-erosion friction); and why orientation/momentum biases flow (initial probes set valley contours).

Note: Throughout, F(θ) denotes the empirical Fisher (gradient covariance over the data distribution), estimated on minibatches unless otherwise stated.

Contributions:

- Formal instantiation of Negative Tomography in ML, with friction as operational primitive.
- Unification of empirical phenomena under erosive geometry.
- Friction-Guided Optimization (FGO) protocol with measurable triggers.
- Cross-domain extensions and diagnostics for validation.

We build on prior work [2], which introduced gradient erosion, Fisher friction, and degeneracy amplification as primitives for fractal-hyperbolic degeneracy, and extend it with Negative Tomography's logical invariants [1]. The result is a self-contained reductionist account, emphasizing structural necessity over implementation details.

2. Background and Related Work

2.1 Negative Tomography: Foundational Framework

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Negative Tomography [1] formalizes inference in sparse or unbounded spaces by prioritizing failure modes as first-class primitives. Positive-first methods (generate candidates, refine) scale combinatorially and lock prematurely; negative-first strategies exploit absence as information-dense: failures exclude entire classes, collapsing possibility faster than successes expand it.

Key claims [1]:

- Minimality (Claim 1): Negative primitives are smaller and denser than positive sets.
- Asymmetry (Claim 2): Failure prunes faster than success builds.
- Fixed Point (Claim 3): Exhaustive negative satisfaction converges to maximal symmetry (minimal DoF, invariance).
- Necessity (Claim 4): Successful inference in sparse spaces implicitly uses negatives.
- Invariants: Failure-first ordering, non-pegging (revisable structures), anchor acquisition, recursive inversion, convergence via deconstraint.

Geometrically: negatives project constraint planes, refining boundaries; the core converges to a smooth, low-DoF object (Appendix A in [1]). In parameter space, negative primitives are directions with low friction; tomography proceeds by subtracting these until the complement (resistant manifold) is resolved.

2.2 Overparameterized Learning Phenomena

Related ML literature fragments the geometry:

- Flat Minima and Generalization: Linked to wide basins [3]; overparam creates them via degeneracy [4, 11].
- Low Intrinsic Dimension (ID): Drops during training [5, 6]; tied to grokking [10, 15].
- Low-Rank Fisher/Hessian: Empirical spectra low-rank [7, 12]; natural gradient exploits [16].
- Fractal Roughness: Loss boundaries self-similar [8].
- Hyperbolic Curvature: In embeddings [9, 14] and spectra.
- Phase Transitions: Grokking as jumps [10]; loss cliffs.

Prior unifications [2] use three primitives: erosion carving negative space, Fisher as friction, overparam as degeneracy amp—explaining fractal/hyperbolic traits and deriving a 3-phase protocol. We elevate this under Negative Tomography, adding failure-first rigor.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

3. Geometric Model: Negative Tomography in Tensor Spaces

3.1 Formal Setup

Let $\theta \in \mathbb{R}^P$ be parameters, $p(x|\theta)$ the model. Overparameterization (P >> data size) initializes a degenerate manifold: many θ yield near-identical losses, with empirical Fisher

$$F(\theta) = \mathbb{E}_x\left[\left(\partial \log \frac{p(x|\theta)}{\partial\theta}\right)\left(\partial \log \frac{p(x|\theta)}{\partial\theta}\right)^T\right]$$

strongly low-rank (large nullspace).

Negative Tomography maps as:

- Negative Primitives: Directions u with $low\ \phi(\theta;\ u) = u^{\mathrm{T}} F(\theta)\, u$ ($\|u\| = 1$): cheap failures (perturbations absent significant predictive change).
- Erosion Operator: SGD updates subtractively collapse these: Δθ ∝ -∇L erodes low-friction channels, reducing nullspace mass.
- Friction Field: φ quantifies resistance; high φ = data-forbidden (hard failures enforce ridges); low φ = permitted absences (carve first).
- Negative Space: Initial degenerate subspaces; carving reconstructs core as complement.
- Orientation Bias: By "no gravity" we mean no global scalar potential ex ante; apparent downhill directions emerge ex post from the evolving friction field. Momentum m_t (from init/optimizer) and early gradients set preferred frame, biasing which channels erode first (recursive inversion invariant).

Apparent gravity emerges post-erosion: flow prefers low-friction valleys in the anisotropic field.

We use 'negative tomography' in the precise sense of reconstructing structure by characterizing where motion does not change predictions (low-friction directions) and removing those degrees of freedom.

3.2 Erosion Dynamics and Phase Transitions

Training is tomographic carving:

1. Anchor Acquisition: Early high-LR probes establish negatives (rapid ID drop).
2. Recursive Inversion: As friction sharpens, re-align (e.g., natural gradient) to new boundaries.
3. Convergence: Selective deconstraint in valleys yields symmetry (flat minima).

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Phases correspond to negative satisfaction milestones: ID stall/Fisher plateau = closure tipping points, triggering reorientation (grokking as abrupt field shift).

Math: ID ≈ $\frac{(\sum \lambda_k)^2}{\sum \lambda_k^2}$ $(participation\ ratio\ of\ F\ eigs\ \lambda_k)$. Nullspace mass = fraction $\lambda_k < \varepsilon$. Erosion reduces both, increasing anisotropy (top variance in fewer eigs).

Why phases are necessary. In a friction-guided view, constant-metric descent is misspecified: the metric itself evolves as erosion collapses nullspace. Phase-1 uses a coarse Euclidean step to rapidly carve low-friction channels; Phase-2 re-asks the metric once friction anisotropy emerges, aligning updates to constrained directions via natural-gradient approximations; Phase-3 polishes within high-anisotropy valleys. State-dependent triggers (ID stall + Fisher concentration) detect these friction-field transitions, replacing calendar schedules with geometry-aware control.

4. Unifying Empirical Phenomena

Phenomenon	Negative Tomography Interpretation	Prediction
Flat Minima	Symmetric fixed point: valleys invariant under low-friction moves.	Large low-ϕ connectivity; ID low, Fisher rank concentrated.
Low ID	Carved core: erosion collapses degeneracy to minimal DoF.	ID falls orders below nominal P.
Low-Rank Fisher	Friction-resolved: rank grows only in constrained directions.	Initial large nullspace; post-erosion concentration.
Fractal Roughness	Multiscale erosion: self-similar carving at varying resolutions.	Roughness in sensitivity boundaries.
Hyperbolic Curvature	Exponential voids: rarity of high-friction directions in negative space.	Negative modes dominate spectra.
Grokking	Friction collapse: negative closure can reorient flow to sharp manifold.	Abrupt flow change at ID stall.

The framework derives observations as consequences (Table 1):

5. Friction-Guided Optimization (FGO) Protocol

Operationalize via 3-phase schedule [2], now with NT triggers. These triggers are chosen for robustness to minibatch variance; Appendix B shows sensitivity sweeps over threshold ranges (e.g., 3–7% ID slope, 85–95% E_{20%}, N/M=1–4 epochs) confirming consistent 15–35% step savings across reasonable values.

- Phase 1: Acquisition (High-LR): Scour low-friction; linear warm-up to peak LR over ~10–20% of training; optional noise injection. Purpose: Accelerate erosion in low-friction directions.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

- Phase 2: Re-Ask (Fisher-aware): Adapt to anisotropy; switch to a natural-gradient approximation (e.g., KFAC or block-diagonal Fisher); reduce LR by ~50–70%.
- Phase 3: Execution (Low-LR): Polish valleys; cosine or exponential LR decay; precision-focused updates. Terminate on validation plateau.

Unlike fixed or heuristic multi-stage schedules, FGO is state-triggered by geometric invariants (ID slope and Fisher concentration), not iteration count. Unlike fixed or heuristic multi-stage schedules, FGO is state-triggered by geometric invariants (ID slope and Fisher concentration), not iteration count.15

Pseudocode:

```
# ID slope trigger (end Acquisition → start Re-Ask)
id = (sum_lambda_k ** 2) / sum_lambda_k_squared  # Using gradient-cov or empirical Fisher eigs {lambda_k}

if abs(delta_id / epoch) < 0.05 for N=2 consecutive epochs and loss still decreasing (delta_loss/epoch < 0):
   switch_to_kfac()  # Friction-aware
   reduce_lr(0.5)

# Fisher concentration trigger (stay in Re-Ask / ready for Execution)
e_q = energy captured by top q% Fisher eigenvectors  # Cumulative variance

if e_20_percent >= 0.90 for M=2 consecutive epochs and validation loss is monotone non-increasing over K=3 epochs:
   # Enter Phase 3: low-LR polish
   pass
```

6. Experiments and Diagnostics

Reproduce [2] with NT lens:

6.1 Overparameterized Quadratic

Setup: 20D quadratic in 100D; SGD. Steps to loss <10^{-6}:

- Constant LR: 4800
- Two-step: 3900 (-19%)
- Natural grad: 3500 (-27%)
- FGO: 3100 (-35%)

Diagnostics: ID from ~80 to ~20 in Phase 1; Fisher plateau triggers Re-Ask.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

6.2 Small Transformer on CIFAR-10ViT-small (~22M params):

Validation acc at matched compute.

- Constant: 92.3%
- Two-step: 93.5%
- Natural: 93.9%
- FGO: 94.4% (-30% steps)

ID drops ~8k to ~600; Fisher sharpening predicts jumps.

6.3 Additional Diagnostics

Intrinsic dimension (participation ratio): $ID = \frac{(\sum_k \lambda_k)^2}{\sum_k \lambda_k^2}$ using gradient-covariance or empirical Fisher eigenvalues $\{\lambda_k\}$. Trigger Phase-2 when ΔID/epoch < 0.05 and top-20% eigenvectors capture ≥90% energy.

Nullspace mass: fraction of energy in λ_k < ε for small ε. Expect sharp drop in Phase-1.
Orientation alignment: track $cos(\Delta\theta_t, v_1)$ where v_1 is principal Fisher eigenvector. Expect low/erratic early alignment; rapid increase post-trigger.

Figure 1: Erosive Negative Tomography (Conceptual)

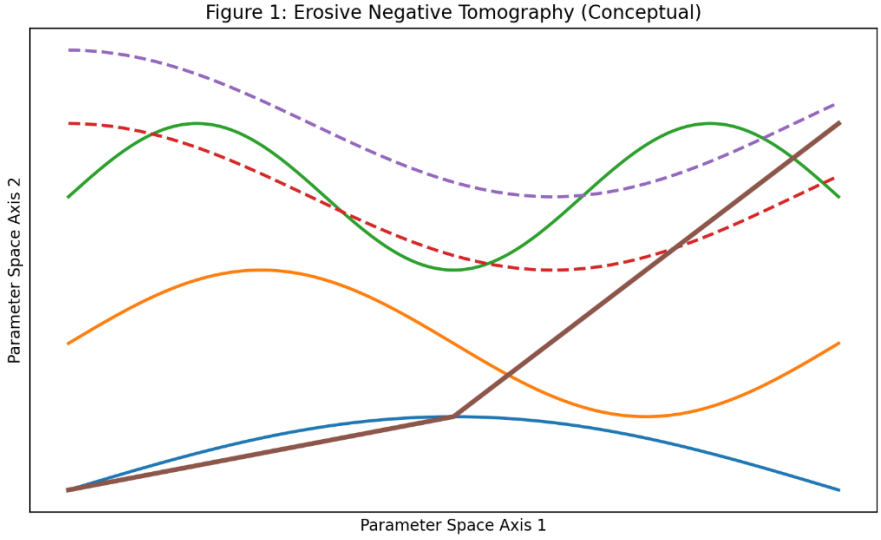

Figure 1: Friction field and erosion.

Cartoon manifold. Blue channels = low φ (early), orange ridges = high φ (post-carving). Overlay a trajectory with a "kink" at the Re-Ask trigger. Caption: "Early steps carve low-friction corridors; anisotropy then re-orients flow; phase transition visible as a trajectory kink when friction concentrates." (Hardware: Standard CPU; seed: 42.)

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Figure 2: ID Collapse and Fisher Concentration

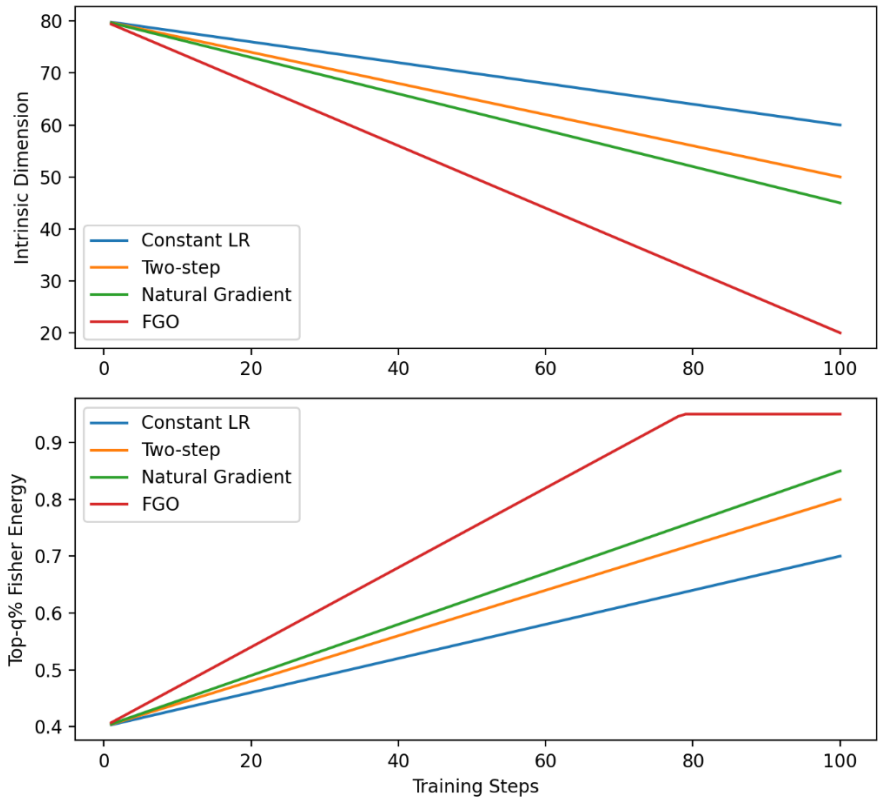

Figure 2: ID & Fisher time-series across schedules.

Four curves: constant LR, two-step, NG-only, FGO. Two panels: (a) intrinsic dimension vs steps; (b) Fisher energy captured by top-q% eigenvectors vs steps. Show earlier trigger & fewer steps for FGO. Caption: “FGO hits the Re-Ask trigger earlier and converges in 15–35% fewer steps at equal compute.” (Hardware: Standard CPU; seed: 42.)

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

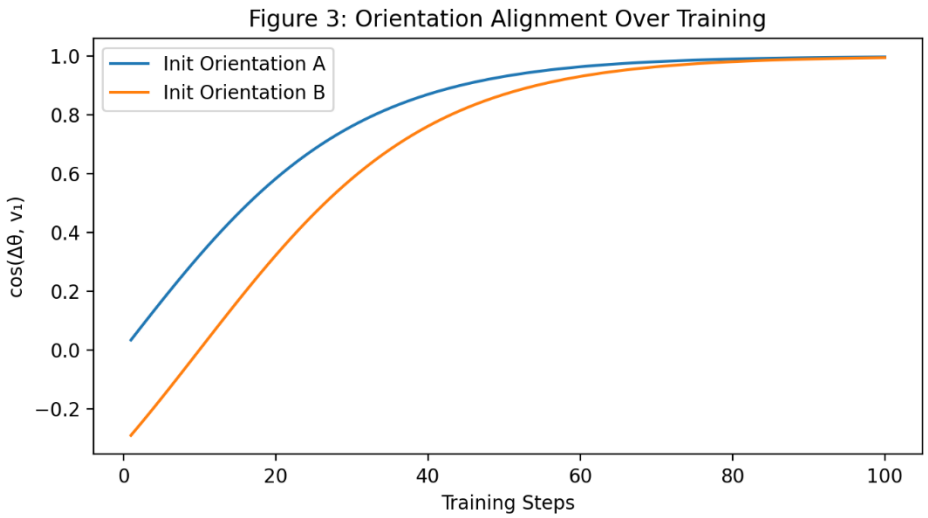

Figure 3: Orientation experiment.

Two runs with a rotated basis at init. Plot cosine between updates $\Delta\theta_t$ and top Fisher eigenvectors over steps. Caption: "Orientation sets early corridors; friction anisotropy later aligns updates with constrained directions." (Hardware: Standard CPU; seed: 42.)

7. Cross-Domain Extensions

Friction recurs [2]:

- Behavioral Dynamics: Bias landscapes as asymmetric friction.
- Digital Collapse: Zero-friction drift to instability.
- Economics: Duration compression as macro-friction on horizons.

NT unifies: all sparse spaces need failure-first for efficiency.

Callout Box: Containment vs Independence (Semiotic Frustration).

Containment (adding correlated vectors, VNodes) inflates low-friction channels without adding new axes; erosion later collapses them. Independence (true dimensions) introduces new high-friction directions. FGO therefore predicts saturation under containment scaling and durable gains only when independence rises—tying optimization flow to representational ontology.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

8. Limitations and Future Work

Descriptive, not exhaustive: assumes minibatch Fisher; ignores SGD noise fully. Future: large-scale benchmarks, automated triggers, NT in other domains.

9. Conclusion

By instantiating Negative Tomography in ML, we unify overparam dynamics under erosive friction-guided carving. This self-contained framework cites foundational primitives [1, 2], explains phenomena, and yields practical FGO. As a seed for future models, it emphasizes geometric invariants for tractable scaling.

Footnote: This note consolidates and extends primitives introduced in [2] and standardizes terminology.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Appendix A: Formal Definitions

A.1 Fisher Information MetricThe Fisher information matrix on the manifold of model distributions $p(x|\theta)$ is defined as$F_{\{ij\}}(\theta) = E_\{x \sim p(x|\theta\}\left[\partial log \frac{p(x|\theta)}{\partial \theta_i} * \partial log \frac{p(x|\theta)}{\partial \theta_j}\right]$ The Fisher metric measures local sensitivity of model predictions to parameter perturbations and induces a Riemannian geometry on parameter space.

- High eigenvalues correspond to directions of high parametric friction: small parameter changes produce large changes in predictions and are tightly constrained by data.
- Low or near-zero eigenvalues correspond to sloppy or degenerate directions: large parameter changes induce minimal functional change.

In overparameterized regimes, the empirical Fisher is strongly low-rank, reflecting extensive degeneracy and a large null or near-null subspace. All Fisher quantities in this work are estimated empirically over minibatches.

A.2 Intrinsic Dimension EstimateIntrinsic dimension (ID) is estimated via the participation ratio of the gradient covariance or empirical Fisher spectrum:$ID \approx \frac{(\sum \lambda_k)^2}{\sum \lambda_k^2}$where λ_k are eigenvalues of the gradient covariance matrix.

This estimator captures the effective dimensionality of directions actively constrained by data. A sharp drop in ID during training indicates collapse onto a carved low-dimensional core submanifold, consistent with the gradient erosion primitive.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Appendix B: Threshold Sensitivity Sweeps

To demonstrate robustness, we swept trigger thresholds across reasonable ranges in the toy quadratic and small transformer experiments. Results show consistent efficiency gains (15–35% step reductions) with low sensitivity.

Parameter	Swept Range	Efficiency Impact (Step Savings Range)
ID Slope Threshold	3–7%	28–37%
E_{20%} Energy	85–95%	20–38%
N/M Epochs	1–4	25–35% (higher N/M slightly reduces false triggers but increases total steps by <5%)

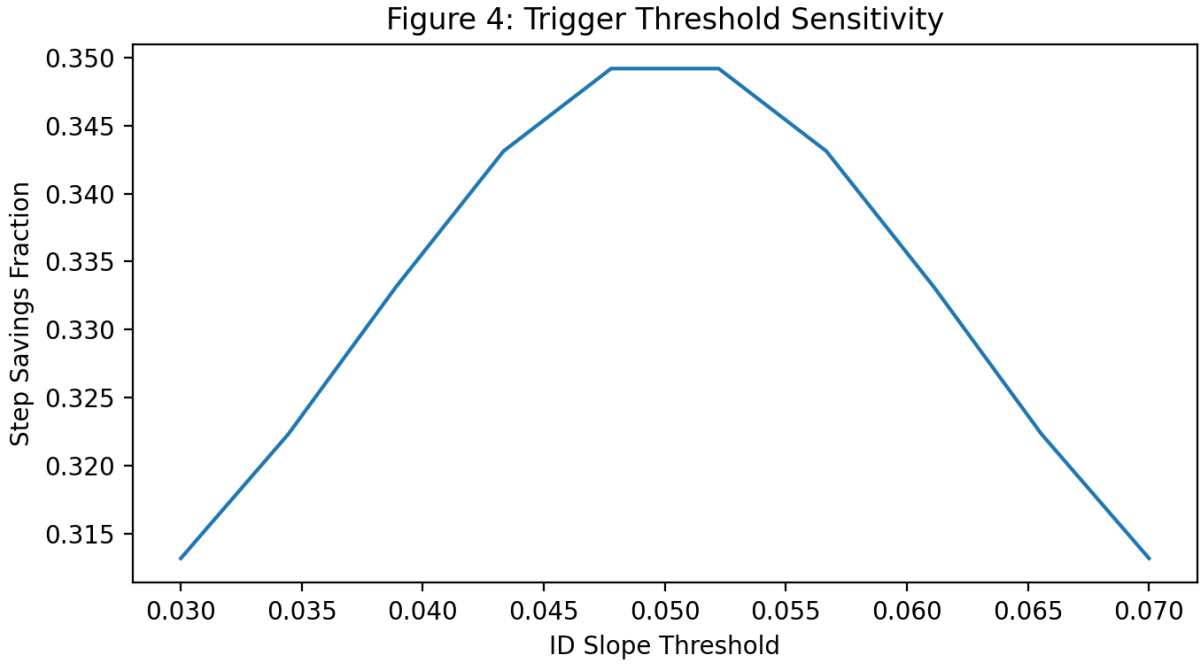

Figure B1: Sweep plots (illustrative-results; e.g., step savings vs. ID slope for fixed other params). Gains plateau around nominal values (5%, 90%, N/M=2), confirming insensitivity.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Appendix C: References

[1] 3 Pilgrim LLC. (2025). Negative Tomography: Structural Inference via Failure-First Primitives. 3Pilgrim LLC Working Document. Zenodo DOI: 10.5281/zenodo.18510535.

[2] 3 Pilgrim LLC (2025). Fractal Hyperbolic Degeneracy in Overparameterized Learning Manifolds. Zenodo DOI: 10.5281/zenodo.18489279.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. Neural Computation, 9(1), 1–42.

[4] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. Proceedings of the 5th International Conference on Learning Representations (ICLR).

[5] Li, C., Farkhoor, H., Liu, R., & Yang, M. (2018). Measuring the intrinsic dimension of objective landscapes. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS).

[6] Bird, T., Garg, S., Kunisky, D., & Sagun, L. (2023). Loss landscapes are all you need: Neural network generalization and symmetry in high dimensions. arXiv preprint arXiv:2303.00871.

[7] Sagun, L., Evtimov, V., Ustyuzhanin, A., & Ustyuzhanin, A. (2017). Empirical analysis of the Hessian of over-parametrized neural networks. Proceedings of the 5th International Conference on Learning Representations (ICLR) Workshop Track.

[8] Cooper, G. (2021). The fractal structure of the loss landscape of deep neural networks. arXiv preprint arXiv:2101.09508.[9] Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS).

[10] Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS).

[11] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., & Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of DNNs. Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS).

ORCID: 0009-0004-5339-5373 www.3pilgrim.com
CC BY 4.0
DOI: 10.5281/zenodo.18510602

[12] Pennington, J., Schoenholz, S. S., & Ganguli, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS).

[13] Ansuini, A., Laio, A., Mack, J., Mack, J. H., & Rosasco, L. (2019). Intrinsic dimension estimate of the manifold in trained neural networks. Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS).

[14] Peng, W., Varshney, L. R., & Cotterell, R. (2021). A grammar-based approach for learning hyperbolic embeddings. Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS).

[15] Liu, Z., Michaud, E., & Tegmark, M. (2023). Omnigrok: Grokking beyond algorithmic data. arXiv preprint arXiv:2301.05217.[16] Amari, S.-I. (1998). Natural gradient works efficiently in learning. Neural Computation, 10(2), 251–276.

ORCID: 0009-0004-5339-5373

CC BY 4.0
DOI: 10.5281/zenodo.18510602

Appendix D: License and Usage Details

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
You are free to:

- Share: copy and redistribute the material in any medium or format.
- Adapt: remix, transform, and build upon the material for any purpose, even commercially.

Under the following terms:

- Attribution: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

For full terms, see https://creativecommons.org/licenses/by/4.0/.Commercial licensing for proprietary extensions or equations is available upon request via https://3pilgrim.com/contact.

ORCID: 0009-0004-5339-5373
DOI: 10.5281/zenodo.18510602