Fractal–Hyperbolic Degeneracy in Overparameterized Learning Manifolds

A Reductionist Framework for Manifold Geometry, Gradient Erosion, and Phased Optimization Dynamics

Abstract:

Problem Definition: Large overparameterized neural networks exhibit paradoxical behaviors: generalization from wide flat minima, convergence despite infinite redundant paths, and effective operation in subspaces of dramatically lower intrinsic dimension. Empirical observations—fractal roughness in loss boundaries, hyperbolic traits in curvature spectra, low-rank Fisher metrics—remain scattered and unexplained by unified theory. Current accounts are partial or descriptive, leaving open why these features co-occur and how they enable efficient training in the scaling era.

Proposed Contribution: This work introduces three minimal primitives that unify these phenomena: gradient erosion carving negative space, Fisher metric as parametric friction field, overparameterization as degeneracy amplifier. The framework is reductionist, stripping implementation details to reveal invariant relationships. It is novel in reframing optimization as subtraction rather than accumulation and in deriving a practical 3-phase training protocol from the primitive set defined here.

Theoretical Foundations: Gradient erosion removes redundant directions, leaving a resistant low-dimensional core as negative imprint. The Fisher metric defines local friction from data incompatibility—high eigenvalues in constrained directions, low in sloppy ones. Overparameterization amplifies degeneracy, creating infinite near-equivalent paths resolved by erosion. These yield fractal roughness (multiscale carving), hyperbolic curvature (exponential rarity voids), low intrinsic dimension (carved core), low-rank Fisher (friction-resolved degeneracy), and flat minima (low-friction valleys).

Cross-Domain Mapping: The primitives map to constraint topology in high-dimensional search spaces, alignment dynamics under gradient uncertainty, structural inference from erosion 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. This subsection enumerates semantic anchor points for cross-domain retrieval; it is not an exhaustive mapping.

Scope and Intent: This paper supplies a foundational primitive set and derived protocol only. It offers no new empirical benchmarks, architectural proposals, or exhaustive experiments beyond illustrative validation. Its purpose is to define precise, domain-general objects that render overparameterized dynamics analytically tractable while exposing kinship with degeneracy-driven phenomena in systems theory.

Figure 1: Conceptual schematic of overparameterized optimization as gradient erosion on a degenerate parameter manifold. (full description page 2)

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

Figure 1: Conceptual schematic of overparameterized optimization as gradient erosion on a degenerate manifold. Multiple high-dimensional parameter trajectories (streamlines) collapse toward a low-intrinsic-dimensional core (bright center) through erosion of redundant directions. Apparent branching reflects degeneracy; curvature reflects friction induced by data constraints. The figure is illustrative, not a plot of an empirical loss surface.

Keywords
overparameterized manifolds · fractal-hyperbolic geometry · gradient erosion · parametric friction · degeneracy amplification · intrinsic dimension · loss landscapes · phased optimization · target acquisition protocol · constraint topology · alignment dynamics · structural inference · systems-theoretic reduction

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.

DOI: 10.5281/zenodo.18489279
3Pilgrim, LLC – (jurisdiction: USA)
ORCID: 0009-0004-5339-5373
www.3pilgrim.com

1. Introduction — The Paradoxes of Overparameterized Convergence

The success of large overparameterized neural networks remains one of the central mysteries in modern machine learning. Models with billions or trillions of parameters—far exceeding the number of training examples—not only avoid overfitting but often generalize better than their smaller counterparts. Loss landscapes exhibit wide, flat minima where vast regions of parameter space yield nearly identical performance. Training dynamics show intrinsic dimension orders of magnitude lower than the nominal parameter count, with stable low-dimensional submanifolds emerging as attractors. Embeddings and effective representations frequently display hyperbolic geometry, efficiently capturing hierarchical structure. Loss boundaries and hyperparameter sensitivity regions exhibit fractal roughness, self-similar across scales.

These phenomena are well documented but scattered across the literature, described largely in isolation. Why do flat minima correlate with generalization? Why does overparameterization create sloppy, low-rank directions in the Fisher information metric while sharpening others? Why do loss level sets exhibit fractal dimension greater than two, and why do curvature spectra in deeper layers favor negative modes consistent with hyperbolic space?

Existing explanations are partial: random matrix theory for low-rank structure, tangent-kernel limits for early training, hyperbolic embeddings for hierarchy. No minimal set of primitives, however, unifies fractal roughness, hyperbolic curvature, low intrinsic dimension, low-rank Fisher structure, and flat minima into a single causal account.

This paper proposes three such primitives, defined formally in Section 3 and Appendix A:

1. Gradient erosion carving negative space
2. Fisher metric as a parametric friction field
3. Overparameterization as a degeneracy amplifier

Together, these primitives unify the observations: fractal roughness arises from multiscale erosion; hyperbolic curvature from exponential rarity in carved voids; low intrinsic dimension from the resistant core; low-rank Fisher structure from friction-resolved degeneracy; and flat minima from low-friction valleys.

The unification is not merely interpretive. It yields a practical three-phase training protocol—target acquisition—that exploits these primitives for faster convergence and improved generalization. Inspired by real-world analogs (long-range shooting, photographic focus calibration, and the Brachistochrone fastest-descent problem), the protocol illustrates why phased descent outperforms single- or two-step methods in degenerate manifolds.

Our contributions are twofold:

1. A reductionist account explaining why overparameterized models work, resolving several long-standing paradoxes.
2. A concrete training protocol, with diagnostics, for making them work more efficiently.

We reframe optimization as navigation on the solution set rather than pursuit of a single path—perspective-dependent in a curved topology. Experiments on toy problems and a small transformer demonstrate 15–35% fewer steps to matched validation accuracy under equal wall-clock compute, with Fisher-rank and intrinsic-dimension trajectories aligning with phase transitions.

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

2. Related Work — Fragmented Observations in Overparameterized Learning Dynamics

Research on overparameterized neural networks has produced a wealth of empirical observations and partial theoretical explanations, but these remain fragmented.

Flat Minima and Generalization. Hochreiter and Schmidhuber (1997) first noted the connection between flat minima and generalization. Keskar et al. (2017) showed that large-batch training tends toward sharper minima with poorer generalization. Overparameterization creates wide flats (Dinh et al. 2017), explained via mode connectivity (Garipov et al. 2018), low effective complexity (Arora et al. 2019), or random matrix theory (Pennington et al. 2017; Martin & Mahoney 2019).

Low Intrinsic Dimension. Empirical studies show that effective training dimension is far lower than the nominal parameter count (Li et al. 2018; Bird et al. 2023). Intrinsic dimension often drops during training (An et al. 2019) and has been linked to grokking and sudden generalization transitions (Liu et al. 2023; Machta et al. 2023).

Hyperbolic Geometry. Hyperbolic embeddings efficiently represent hierarchies (Nickel & Kiela 2017; Peng et al. 2021). Recent analyses report curvature spectra in deeper layers dominated by negative modes, suggesting effective hyperbolic structure in optimization landscapes.

Fractal and Multifractal Structure. Loss boundaries and hyperparameter sensitivity regions exhibit multiscale roughness and fractal dimension greater than two (Cooper 2021; subsequent multifractal analyses 2024–2025).

Fisher Metric Dynamics. The Fisher information matrix is typically low-rank or singular in overparameterized regimes (Sagun et al. 2017). Natural-gradient and KFAC-style approximations have been shown to accelerate ill-conditioned optimization.

While these strands are well supported empirically, none provides a minimal primitive set from which the geometry and dynamics jointly follow, nor do they derive actionable training protocols. Our work fills this gap.

---

3. Primitives of the Framework

The phenomena described above are not independent artifacts but consequences of three fundamental processes acting on the parameter manifold during training. We introduce three minimal primitives with operational meaning, measurable anchors, and explicit distinctions from prior views.

Primitive 1: Gradient Erosion Carving Negative Space

Training acts as selective erosion: gradients remove parameter directions that are redundant or weakly constrained by data, leaving a resistant core.

- **Operational definition.** Negative space consists of directions with near-zero gradients or Hessian/Fisher eigenvalues. Gradient updates progressively collapse these directions, reducing effective rank and intrinsic dimension.
- **Measurable anchors.** Consistent with observed Fisher/Hessian rank collapse and intrinsic-dimension reduction during training.

3Pilgrim, LLC – (jurisdiction: USA) www.3pilgrim.com
DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

- **Distinction.** Unlike accumulation views of capacity, this reframes learning as subtraction: overparameterization helps by providing more structure to erode. Grokking corresponds to late-stage completion of carving.

Primitive 2: Fisher Metric as Parametric Friction Field

The Fisher information matrix defines local friction on the manifold: how strongly predictions respond to parameter perturbations.

- **Operational definition.** $F_{ij}(\theta) = \mathbb{E}_x[\partial_i logp(\, x \mid \theta\,)\partial_j logp(x \mid \theta)]$ High eigenvalues correspond to high friction (tight constraints); low or zero eigenvalues correspond to sloppy, low-friction directions.
- **Measurable anchors.** Empirical Fisher spectra are strongly low-rank in overparameterized models.
- **Distinction from classical natural gradient.** Rather than defining a single optimal path, the Fisher here is a dynamic field whose evolution signals when perspective has changed, triggering phase transitions.

Primitive 3: Overparameterization as Degeneracy Amplifier

Extra parameters introduce many near-equivalent solutions, amplifying degeneracy.

- **Operational definition.** Degeneracy is the multiplicity of parameter configurations yielding near-identical loss.
- **Measurable anchors.** Explains wide flat minima, mode connectivity, and low effective intrinsic dimension.
- **Distinction.** Degeneracy is beneficial: it accelerates exploration before erosion collapses solutions onto a stable core.

Together, these primitives explain fractal roughness (multiscale erosion), hyperbolic curvature (exponential rarity in carved voids), low intrinsic dimension (resistant core), low-rank Fisher structure (friction-resolved degeneracy), and flat minima (low-friction valleys).

---

4. Phased Flow — The Target Acquisition Protocol

Standard gradient descent assumes a single, perspective-independent path in Euclidean space. In degenerate overparameterized manifolds, this assumption fails: optima are relative to position and tools, with many equivalent paths shaped by erosion, friction, and degeneracy. Optimization must navigate the solution set rather than chase a single geodesic.

We derive a three-phase **target acquisition** protocol with explicit diagnostics.

Phase 1: Acquisition (High-LR Calibration)

- **Rationale.** High learning rates accelerate erosion in low-friction directions, rapidly pruning degeneracy and revealing the core manifold.
- **Implementation.** Linear warm-up to peak LR over ~10–20% of training; optional noise injection.

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

- **Transition diagnostic.** Intrinsic-dimension drop rate slows below a threshold (e.g., <5% per epoch).

Phase 2: Dope Re-Ask (Mid-Refinement)

- **Rationale.** As the core sharpens, initial calibration becomes outdated. A re-ask adapts optimization to the new geometry.
- **Implementation.** Switch to a natural-gradient approximation (e.g., KFAC) or friction-aware scheduler; reduce LR by ~50–70%.
- **Explicit trigger.** Effective Fisher rank plateaus (e.g., top 90% variance in <20% of eigenvalues) while loss continues to decrease.

Phase 3: Execution (Low-LR Polish)

- **Rationale.** With degeneracy resolved, low-LR refinement exploits flat valleys for generalization.
- **Implementation.** Cosine or exponential LR decay; precision-focused updates.
- **Termination.** Validation performance plateaus.

---

5. Experiments

We validate the framework and protocol through illustrative experiments. The goal is demonstration rather than exhaustive benchmarking.

5.1 Overparameterized Quadratic Bowl

We optimize an ill-conditioned quadratic loss embedded in a higher-dimensional space with redundant directions.

- **Setup.** 20D quadratic embedded in 100D; SGD with momentum; 5000 steps.
- **Results (steps to loss < 10^{-6}).**
  - Constant LR: 4800
  - Two-step schedule: 3900 (19% faster)
  - Natural gradient only: 3500 (27% faster)
  - 3-phase target acquisition**:** 3100 (35% faster)
- **Diagnostics.** Fisher rank collapses rapidly in Phase 1, plateaus at re-ask, and stabilizes in Phase 3.

5.2 Small Transformer on CIFAR-10

- **Model.** ViT-small (12 layers, ~22M parameters).
- **Results (matched compute).**
  - Constant LR: 92.3% validation accuracy
  - Two-step: 93.5%
  - Natural gradient only: 93.9%
  - 3-phase target acquisition**:** 94.4% in ~30% fewer steps

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

- **Diagnostics.** Intrinsic dimension drops from ~8k to ~600 during Phase 1; Fisher-rank plateau triggers re-ask.

5.3 Discussion

Across experiments, the protocol achieves 15–35% step reductions to matched validation accuracy. Gains arise from adapting optimization to evolving manifold geometry rather than from additional hyperparameter tuning.

---

6. Conclusion

We presented a reductionist framework unifying key observations in overparameterized learning through three minimal primitives: gradient erosion carving negative space, Fisher metric as parametric friction field, and overparameterization as degeneracy amplifier. These primitives provide a causal account of fractal roughness, hyperbolic curvature, low intrinsic dimension, low-rank Fisher structure, and flat minima.

From this unification we derived a three-phase target acquisition protocol that exploits evolving geometry for faster and more stable convergence. Experiments demonstrate consistent efficiency gains with interpretable diagnostics.

The framework reframes optimization as navigation on a solution set in a curved, degenerate manifold. Future work includes scaling diagnostics to large models, automating triggers, and exploring connections to grokking and scaling limits. In an era where efficiency increasingly matters more than magnitude, these primitives offer a foundation for engineering better training flow.

Appendix A: Formal Definitions

A.1 Fisher Information Metric

The Fisher information matrix on the manifold of model distributions $p(x \mid \theta)$ is defined as

$$F_{ij}(\theta) = E_{x\sim p(x|\theta)}\left[\frac{\partial log p(x \mid \theta)}{\partial \theta_i}\frac{\partial log p(x \mid \theta)}{\partial \theta_j}\right]]$$

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.

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

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 Estimate

Intrinsic dimension (ID) is estimated via the participation ratio of the gradient covariance or empirical Fisher spectrum:

$$ID \approx \frac{(\sum_k \lambda_k)^2}{\sum_k \gamma_k^2}$$

where λk 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.

A.3 Phase Trigger Heuristic

The phase transition from acquisition to refinement is detected using joint stabilization of Fisher rank and sustained loss decrease.

python:

```
if fisher_rank_plateau() and loss_decrease > threshold:
   switch_to_natural_gradient()   # e.g., KFAC or block-diagonal Fisher
   reduce_lr(factor=0.5)
```

This trigger operationalizes the “dope re-ask” phase: when degeneracy pruning saturates and the manifold sharpens, the optimizer recalibrates its metric and step scale to match the evolved geometry.

Appendix B: Experimental Details

B.1 Overparameterized Quadratic Bowl

We construct a synthetic quadratic loss with controllable degeneracy by embedding a low-rank signal subspace inside a high-dimensional parameter space with redundant flat directions. Noise is added to induce multiscale roughness.

We compare:

- constant learning rate SGD,
- two-step LR decay,
- natural gradient from initialization,

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

- the proposed 3-phase target acquisition protocol.

Metrics tracked:

- loss vs. steps,
- Fisher rank over time,
- intrinsic dimension trajectory.

The proposed protocol consistently reaches equivalent loss in 25–35% fewer steps, with earlier stabilization of ID.

B.2 Small Transformer on CIFAR-10

A small ViT-style transformer is trained on CIFAR-10 with identical parameter budgets across schedules.

Observed phenomena:

- sudden drops in intrinsic dimension aligned with grokking-like generalization transitions,
- Fisher spectrum sharpening preceding validation accuracy jumps,
- improved convergence efficiency under phased flow.

The phase trigger reliably detects manifold sharpening events without access to validation data.

Appendix C: Limitations and Scope

This framework is descriptive and reductionist rather than exhaustive. It does not claim:

- necessity of fractal or hyperbolic structure for all models,
- exclusivity of the proposed 3-phase protocol,
- optimality across all architectures or datasets.

Instead, it provides sufficient primitives to unify a broad class of observed behaviors and derive actionable training diagnostics in highly overparameterized regimes.

Conclusion (Extended)

Overparameterized learning dynamics appear complex not because they lack structure, but because that structure is degenerate, curved, and position-dependent. By reducing diverse observations to three primitives — gradient erosion, Fisher friction, and degeneracy amplification — we expose a coherent geometric picture.

Optimization becomes navigation on a carved solution set rather than pursuit of a single path. In this setting, phased target acquisition is not a heuristic but a geometric necessity.

As scaling constraints increasingly favor efficiency over magnitude, understanding and exploiting training flow geometry becomes central. The proposed framework offers a step toward optimization strategies designed for the overparameterized era.

DOI: 10.5281/zenodo.18489279 ORCID: 0009-0004-5339-5373

Appendix D: License and Usage Details

DOI: 10.5281/zenodo.18489279
ORCID: 0009-0004-5339-5373