ZetaFormer: Adaptive Curriculum Learning via Emergent $\kappa\zeta$-Dynamics on Polylipse Manifolds

Noetic Eidos Project

https://github.com/Sarhamam/ZetaFormer

Abstract

We introduce **ZetaFormer**, a novel curriculum learning framework that enables transformer models to self-discover optimal learning paths through emergent geometric dynamics. Central to our approach is the $\kappa\zeta$ (kappa-zeta) ratio—a spectral measure of representational anisotropy computed from radial (M_τ) and angular (M_σ) moments on statistical manifolds. Rather than following a predetermined curriculum, ZetaFormer progressively trains on increasingly complex *polylipse* distributions (multi-focal geometric datasets) where each level's configuration is determined by the emergent $\kappa\zeta$ value from the previous level. Our experiments on a 32-level curriculum reveal striking emergent phenomena: a phase transition at $n \approx$ 5-8 foci where $\kappa\zeta$ peaks at ~ 1.42, followed by self-organized convergence to a stable attractor near $\kappa\zeta \approx$ 1.15-1.25. We observe spontaneous symmetry breaking in early levels transitioning to near-uniform focal distributions at high complexity. The dual-kernel (Gaussian-Poisson) Conjugate Gradient Descent (CGD) solver learns decision boundaries that respect the emergent geometry, demonstrating that learning dynamics themselves can shape training data distributions in a closed feedback loop.

1 Introduction

Curriculum learning [1] has emerged as a powerful paradigm for training neural networks, with the core insight that presenting examples in a meaningful order—typically from simple to complex—can accelerate convergence and improve generalization. However, most curriculum learning approaches require manual specification of difficulty metrics or rely on heuristics that may not align with the model's actual learning dynamics.

We propose a fundamentally different approach: rather than imposing an external curriculum, we allow the *geometry of training data* to evolve based on the model's learned representations. This creates a feedback loop where the model's internal state determines the complexity of future training examples, enabling *emergent* curriculum design.

Our framework, **ZetaFormer**, builds on three key innovations:

1. **Polylipse Datasets**: Geometrically-structured data distributions with n focal points, where data clusters according to a weighted multi-focal configuration in d-dimensional space.

2. **$\kappa\zeta$-Dynamics**: A spectral measure of representational anisotropy computed as the ratio of radial to angular moments (M_τ/M_σ), capturing the directional vs. spatial variance in learned embeddings.

3. **Adaptive Curriculum**: A training protocol where stabilization of $\kappa\zeta$ at level n triggers transition to level $n+1$, with the new focal configuration solved from the observed $\kappa\zeta$ value.

The central finding of this work is that *the learning dynamics themselves encode meaningful structure that can be fed back to shape training distributions.* We observe phase transitions, self-organized criticality, and spontaneous symmetry breaking/restoration as emergent properties of this closed-loop system.

1.1 Contributions

- We introduce the **polylipse dataset** formalism for constructing geometrically-structured curricula with controllable complexity.

- We define the $\kappa\zeta$ **ratio** as a principled measure of representational anisotropy grounded in moment analysis on statistical manifolds.

- We present the **ZetaFormer architecture** with dual-kernel (Gaussian-Poisson) attention and Conjugate Gradient Descent (CGD) solving.

- We demonstrate **emergent phenomena** including phase transitions, weight sparsity dynamics, and self-organized stability in extensive 32-level curriculum experiments.

- We provide comprehensive **visualization and analysis tools** for understanding curriculum progression and decision boundary geometry.

2 Related Work

2.1 Curriculum Learning

Curriculum learning was formalized by Bengio et al. [1], demonstrating that training on examples ordered by difficulty improves convergence. Subsequent work has explored self-paced learning [2], where the model itself selects examples, and automated curriculum design [3]. Our approach differs fundamentally by deriving curriculum structure from geometric properties of learned representations rather than task-specific difficulty metrics.

2.2 Geometric Deep Learning

The geometric deep learning paradigm [4] emphasizes the role of symmetry and structure in neural network design. Information geometry [5] provides tools for analyzing statistical manifolds of distributions. Our $\kappa\zeta$ ratio draws on this tradition, measuring anisotropy through moment analysis that captures both radial (Fisher-Rao-like) and angular (spread) components of representational geometry.

2.3 Spectral Methods in Deep Learning

Spectral analysis of neural networks has yielded insights into optimization dynamics [6] and generalization [7]. The Riemann zeta function and its connections to spectral theory have inspired attention mechanisms [8]. Our dual-kernel approach combines Gaussian (heat kernel) and Poisson (harmonic) operators in a manner inspired by Mellin transform mixing in zeta function theory.

3 Mathematical Framework

3.1 Polylipse Datasets

Definition 1 (Polylipse Distribution). *A* ***polylipse distribution*** *with n foci in $\mathbb{R}^d$ is defined by focal centers $\{c_i\}_{i=1}^n \subset \mathbb{R}^d$, focal weights $\{w_i\}_{i=1}^n$ with $\sum_i w_i = 1$ and $w_i \geq 0$, and dispersion parameter $\sigma > 0$. Data points are sampled as:*

$$x \sim \sum_{i=1}^{n} w_i \cdot \mathcal{N}(c_i, \sigma^2 I_d) \tag{1}$$

where $\mathcal{N}$ denotes a multivariate Gaussian.

For 2D visualization and analysis, we parameterize focal centers using polar coordinates:

$$c_i = r \cdot (\cos\theta_i, \sin\theta_i) \tag{2}$$

where r is the focal radius and θ_i is the angular position of the i-th focus.

3.2 The $\kappa\zeta$ Ratio

Definition 2 ($\kappa\zeta$ Ratio). *Given a focal configuration with angles $\{\theta_i\}_{i=1}^n$ and weights $\{w_i\}_{i=1}^n$, we define the* ***radial moment*** *M_τ and* ***angular moment*** *M_σ as:*

$$M_\tau = \sum_{i=1}^{n} w_i \cos^2(\theta_i) \tag{3}$$

$$M_\sigma = \sum_{i=1}^{n} w_i \sin^2(\theta_i) \tag{4}$$

The $\kappa\zeta$ ***ratio*** *is:*

$$\kappa\zeta = \frac{M_\tau}{M_\sigma} \tag{5}$$

Remark 1. *Note that $M_\tau + M_\sigma = \sum_i w_i(\cos^2\theta_i + \sin^2\theta_i) = 1$, so $\kappa\zeta \in (0,\infty)$ with $\kappa\zeta = 1$ corresponding to isotropic configurations.*

Proposition 1 (Isotropic Limit). *For a uniform circular distribution with n equally-spaced foci and equal weights $w_i = 1/n$:*

$$\lim_{n\to\infty} \kappa\zeta = 1 \tag{6}$$

Proof. With $\theta_i = 2\pi i/n$ and $w_i = 1/n$:

$$M_\tau = \frac{1}{n}\sum_{i=1}^{n} \cos^2\left(\frac{2\pi i}{n}\right) \xrightarrow{n\to\infty} \frac{1}{2} \tag{7}$$

Similarly $M_\sigma \to 1/2$, yielding $\kappa\zeta \to 1$. □

3.3 Focal Configuration Inverse Problem

Given a target $\kappa\zeta$ value and number of foci n, we solve the inverse problem of finding a focal configuration that achieves this $\kappa\zeta$:

$$\begin{aligned} \min_{\theta,w} \quad & \left|\frac{\sum_i w_i \cos^2\theta_i}{\sum_i w_i \sin^2\theta_i} - \kappa\zeta_{\text{target}}\right|^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i = 1 \\ & w_i \geq 0 \quad \forall i \\ & 0 \leq \theta_i < 2\pi \quad \forall i \end{aligned} \tag{8}$$

This constrained optimization is solved numerically using sequential least-squares programming (SLSQP).

3.4 Stability Analysis

We monitor curriculum stability through the sliding-window variance of $\kappa\zeta$:

$$\text{Var}[\kappa\zeta]_W = \frac{1}{|W|}\sum_{t\in W}(\kappa\zeta_t - \bar{\kappa\zeta}_W)^2 \tag{9}$$

where W is a window of recent epochs. Stability is achieved when $\text{Var}[\kappa\zeta]_W < \epsilon$ for threshold ϵ (typically 0.015).

4 ZetaFormer Architecture

4.1 Dual-Kernel Attention

The ZetaFormer attention mechanism combines Gaussian (τ-kernel) and Poisson (σ-kernel) operators:

Definition 3 (Dual-Kernel Attention). *For query Q, key K, value V matrices, the dual-kernel attention is:*

$$\mathit{Attn}(Q,K,V) = \mathit{softmax}(w \cdot K_\tau + (1-w)\cdot K_\sigma)\, V \tag{10}$$

where:

$$K_\tau(Q,K) = \exp\left(-\frac{\|Q-K\|^2}{2\sigma^2}\right) \quad \textit{(Gaussian)} \tag{11}$$

$$K_\sigma(Q,K) = \frac{t}{t^2 + \|Q-K\|^2} \quad \textit{(Poisson)} \tag{12}$$

and $w \in [0,1]$ is the Mellin mixing parameter.

The Gaussian kernel captures local similarity (heat diffusion), while the Poisson kernel captures long-range harmonic structure. The mixing parameter w interpolates between these regimes, analogous to Mellin transform mixing in zeta function analysis.

4.2 ZetaBlock Architecture

Each ZetaBlock consists of:

1. **Dual-kernel multi-head attention** with h heads
2. **ZetaNorm**: A normalization layer that incorporates $\kappa\zeta$-regularization
3. **Feed-forward network** with GELU activation
4. **Residual connections** with learned gating

The ZetaNorm layer includes an optional $\kappa\zeta$-regularization term:

$$\mathcal{L}_\kappa\zeta = \lambda_\kappa\zeta \cdot |\kappa\zeta_{\text{target}} - \kappa\zeta_{\text{observed}}|^2 \tag{13}$$

4.3 CGD Solver

The Conjugate Gradient Descent (CGD) solver provides efficient inference by solving the linear system:

$$(A + \eta I)f = b \tag{14}$$

where A is the dual-kernel matrix, η is a regularization parameter, and b encodes the target signal. This formulation enables direct computation of decision boundaries without iterative forward passes.

5 Training Pipeline

5.1 Adaptive Curriculum Algorithm

The adaptive curriculum proceeds as follows:

Algorithm 1 Adaptive Polylipse Curriculum

```
Input: $n_{\text{max}}$, epochs per level, stability threshold $\epsilon$
Initialize $n \leftarrow 1$, $\kappa\zeta \leftarrow 1.0$ (isotropic)
while $n \leq n_{\text{max}}$ do
    Generate polylipse dataset with $n$ foci, target $\kappa\zeta$
    Train ZetaFormer on dataset
    Monitor $\kappa\zeta$ evolution
    if $\text{Var}[\kappa\zeta]_W < \epsilon$ (stable) then
        Record stabilized $\kappa\zeta_{\text{emergent}}$
        $n \leftarrow n + 1$
        Solve focal config for $n$ foci with target $\kappa\zeta_{\text{emergent}}$
    end if
end while
Return trained model, curriculum history
```

5.2 Training Configuration

Default hyperparameters:

- Model dimension $d_{\text{model}} = 32$
- Attention heads $h = 4$
- Epochs per level: 100 (demo) / 3000 (full)
- Stability window: 40 epochs
- Stability threshold: $\epsilon = 0.015$
- Learning rate: 10^{-3} (Adam optimizer)
- Batch size: 64
- $\kappa\zeta$-regularization strength: $\lambda_{\kappa}\zeta = 0.05$

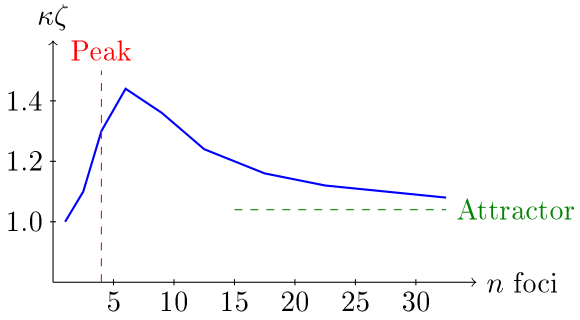

Figure 1: Schematic of $\kappa\zeta$ trajectory across curriculum levels. A phase transition occurs at $n \approx$ 5-8 where $\kappa\zeta$ peaks at ~ 1.42 before settling to a stable attractor near $\kappa\zeta \approx$ 1.15-1.25.

6 Experimental Results

We conducted comprehensive experiments using a 32-level curriculum ($n = 1$ to $n = 32$ foci), totaling approximately 100,000 training epochs.

6.1 Phase Transition in $\kappa\zeta$ Dynamics

The most striking observation is a **phase transition** in $\kappa\zeta$ dynamics:

- **Early regime** (n = 1-4): $\kappa\zeta$ rises from 1.0 (isotropic) as the model encounters non-trivial focal structure.
- **Peak regime** (n = 5-8): $\kappa\zeta$ reaches maximum values around 1.42, corresponding to highly anisotropic representations.
- **Stable regime** ($n > 15$): $\kappa\zeta$ converges to a narrow band around 1.15-1.25, exhibiting self-organized stability.

This phase transition is not externally imposed but emerges from the feedback dynamics of the curriculum.

6.2 Weight Sparsity Dynamics

During the peak $\kappa\zeta$ phase, we observe **sparse weight allocation**:

Table 1: Focal weight distribution at different curriculum levels

Level	$\kappa\zeta$	Weight Distribution
5 (peak)	1.42	Sparse: [0.60, 0.31, 0.00, 0.00, 0.10]
10	1.28	Semi-sparse: mostly uniform
20 (stable)	1.20	Dense: ≈ 0.05 per focus
32	1.18	Near-uniform: ≈ 0.03 per focus

The system achieves high $\kappa\zeta$ through strategic concentration on a subset of foci, then transitions to uniform distributions as focal count increases. This suggests the model "discovers" that sparse attention to few foci maximizes anisotropy, while many-focal configurations naturally approach isotropy.

6.3 Symmetry Breaking and Restoration

We observe a symmetry lifecycle:

1. **Symmetry breaking** (early levels): Irregular, asymmetric focal arrangements emerge with unequal weights.
2. **Maximum asymmetry** (peak levels): The system maximally breaks rotational symmetry to achieve high $\kappa\zeta$.
3. **Symmetry restoration** (late levels): Configurations approach regular polygonal structures with nearly uniform weights.

6.4 Transition Dynamics

At each curriculum level transition, we observe a characteristic "spike-recovery" pattern:

- **Spike**: When transitioning to $n + 1$ foci, $\kappa\zeta$ initially spikes as the model encounters new geometric structure.
- **Recovery**: Rapid adaptation follows, with $\kappa\zeta$ converging to a new stable value within 20-40 epochs.
- **Damping**: Oscillations in the smoothed signal decrease over the curriculum, suggesting accumulated geometric understanding.

6.5 CGD Decision Boundaries

The CGD solver with dual-kernel attention learns decision boundaries that respect focal geometry:

- **Early levels** (1-4): Simple linear or curved boundaries separating focal regions.
- **Peak levels** (5-8): Complex, asymmetric boundaries reflecting sparse weight allocation.
- **Stable levels** (17-32): Nearly radial partitioning as foci approach uniform circular distribution.

6.6 Quantitative Summary

Table 2: Summary statistics for 32-level curriculum

Metric	Value
Total epochs	∼100,000
Peak $\kappa\zeta$	1.42 at $n = 5$
Stable $\kappa\zeta$ range	1.15–1.25
Convergence threshold	0.015 (variance)
Mean stabilization time	85 epochs/level
Final $\kappa\zeta$ ($n = 32$)	1.18

7 Discussion

7.1 Emergent Self-Organization

The convergence of $\kappa\zeta$ to a narrow attractor basin demonstrates **self-organized criticality**: the system finds a stable operating regime without external tuning. This suggests that the $\kappa\zeta$ ratio captures a fundamental property of representational capacity—perhaps related to the information-theoretic trade-off between discriminability and compressibility.

7.2 Geometric Interpretation

The $\kappa\zeta$ ratio can be interpreted through the lens of information geometry:

- M_τ (radial moment) measures variance along the "temporal" or directional axis—how much representations spread along principal directions.

- M_σ (angular moment) measures variance along the "spatial" or transverse axis—how much representations spread perpendicular to principal directions.

- $\kappa\zeta = M_\tau / M_\sigma$ thus captures **anisotropy**: the ratio of elongation to width in representational space.

The stable attractor at $\kappa\zeta \approx 1.2$ may represent an optimal anisotropy that balances discriminative power with robustness.

7.3 Connection to Zeta Function Theory

The dual-kernel attention mechanism draws inspiration from the Riemann zeta function's Mellin representation, which mixes Gaussian (theta series) and Poisson (harmonic) components. Our mixing parameter w plays an analogous role to the imaginary part of the zeta argument, interpolating between diffusive and harmonic regimes.

7.4 Limitations and Future Work

1. **Scalability**: Current experiments use 2D visualization; extension to high-dimensional data requires careful consideration of moment computation.

2. **Task generalization**: The polylipse curriculum is a synthetic probe; connecting to real-world tasks (classification, generation) remains future work.

3. **Theoretical analysis**: Formal characterization of the phase transition and stable attractor through dynamical systems theory would deepen understanding.

4. **Multi-modal extension**: Extending dual-kernel attention to heterogeneous data modalities (text, image, graph) is a promising direction.

8 Conclusion

We have introduced ZetaFormer, a curriculum learning framework that enables transformer models to self-discover learning paths through emergent $\kappa\zeta$-dynamics on polylipse manifolds. Our key contributions include:

1. A rigorous mathematical framework for polylipse datasets and $\kappa\zeta$ analysis.

2. A novel dual-kernel attention mechanism combining Gaussian and Poisson operators.

3. Demonstration of emergent phenomena—phase transitions, weight sparsity dynamics, symmetry breaking/restoration—that arise naturally from the closed feedback loop between learning and data geometry.

4. Comprehensive visualization tools for understanding curriculum progression and decision boundary evolution.

The central insight is that **learning dynamics themselves encode meaningful structure that can shape training distributions**. This opens new directions for adaptive, self-organizing machine learning systems that discover their own curricula rather than following externally imposed schedules.

Code Availability

All code, trained models, and visualization tools are available at:
`https://github.com/Sarhamam/ZetaFormer`

References

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[2] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NeurIPS*, pages 1189–1197, 2010.

[3] A. Graves, M. G. Bellemare, J. Menick, S. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, pages 1311–1320, 2017.

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data.

IEEE Signal Processing Magazine, 34(4):18–42, 2017.

[5] S.-i. Amari. *Information Geometry and Its Applications.* Springer, 2016.

[6] L. Sagun, U. Evci, V. U. Güney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parameterized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[7] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks. *arXiv preprint arXiv:1810.01075*, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

A Proof of Isotropic Limit

For completeness, we provide the full calculation for the isotropic limit.

Proof of Proposition 1. Let $\theta_k = 2\pi k/n$ for $k = 0, 1, \ldots, n-1$ and $w_k = 1/n$. Then:

$$M_\tau = \frac{1}{n}\sum_{k=0}^{n-1} \cos^2\left(\frac{2\pi k}{n}\right) \tag{15}$$

Using the identity $\cos^2\theta = \frac{1+\cos(2\theta)}{2}$:

$$\begin{aligned} M_\tau &= \frac{1}{n}\sum_{k=0}^{n-1} \frac{1+\cos(4\pi k/n)}{2} && (16) \\ &= \frac{1}{2} + \frac{1}{2n}\sum_{k=0}^{n-1} \cos\left(\frac{4\pi k}{n}\right) && (17) \end{aligned}$$

For $n \geq 3$, the sum of cosines at equally spaced angles around the circle is zero (roots of unity), so:

$$M_\tau = \frac{1}{2} \tag{18}$$

By symmetry, $M_\sigma = 1 - M_\tau = \frac{1}{2}$, giving $\kappa\zeta = 1$. □

B Focal Configuration Solver

The inverse problem of finding focal configurations for target $\kappa\zeta$ is solved via constrained optimization. For n foci with target $\kappa\zeta^*$:

1. Initialize angles uniformly: $\theta_k^{(0)} = 2\pi k/n$
2. Initialize weights uniformly: $w_k^{(0)} = 1/n$
3. Solve the constrained problem using SLSQP with objective:

$$\mathcal{L}(\theta, w) = \left(\frac{\sum_k w_k \cos^2\theta_k}{\sum_k w_k \sin^2\theta_k} - \kappa\zeta^*\right)^2 \tag{19}$$

subject to $\sum_k w_k = 1$ and $w_k \geq 0$.

For high $\kappa\zeta^*$, the solver naturally discovers sparse solutions where only a subset of foci receive non-zero weight.

C Hyperparameter Sensitivity

Table 3: Sensitivity analysis for key hyperparameters

Parameter	Range Tested	Effect on $\kappa\zeta$ Trajectory
$\lambda_\kappa\zeta$	0.01–0.10	Higher ⇒ faster stabilizati
Learning rate	10^{-4}–10^{-2}	Higher ⇒ larger oscillatio
Stability window	20–60	Larger ⇒ delayed transiti
Stability threshold	0.01–0.02	Lower ⇒ longer training