Replicator Dynamics in Neural Networks: Geometric Foundations of Softmax and Attention

Christopher Rae Lee-Jenkins *

August 16, 2025

Abstract

We develop a geometric account of softmax as a principal bundle over the probability simplex and project logit flows, via the softmax Jacobian, to replicator dynamics on the simplex. The final head of large language model induces a logit field whose softmax projection yields a replicator evolution of token probabilities. Extended to the whole transformer, attention, gating, and the output head function as stacked evolutionary games that together constitute a multiscale competency architecture, in the sense of Levin.

1 Softmax Geometry

The *softmax map*

$$\sigma : \mathbb{R}^n \to \Delta^{n-1}_{>0}, \qquad \sigma_i(z) = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}},$$

projects *logits* z to the interior of the probability simplex Δ^{n-1}. Shifting logits uniformly leaves σ invariant: $\sigma(z + c\mathbf{1}) = \sigma(z)$. Thus σ is a (trivial, principal) bundle map whose fibers are affine lines $z + \mathbb{R}\mathbf{1}$, with gauge group $(\mathbb{R}, +)$ acting by uniform shifts. In exponential coordinates $x = \exp(z) \in \mathbb{R}^n_{>0}$, fibers are rays $\{\lambda x : \lambda > 0\}$ with projection $\pi(x) = x/(\mathbf{1}^\top x)$.

Write $S(z) = \sum_j e^{z_j}$. Then

$$\frac{\partial \sigma_i}{\partial z_k} = \frac{\partial}{\partial z_k}\Big(\frac{e^{z_i}}{S}\Big) = \frac{\delta_{ik} e^{z_i} S - e^{z_i} e^{z_k}}{S^2} = \frac{e^{z_i}}{S}\Big(\delta_{ik} - \frac{e^{z_k}}{S}\Big) = p_i(\delta_{ik} - p_k).$$

Thus the Jacobian $J(p) \in \mathbb{R}^{n\times n}$ has entries

$$J_{ik}(p) = \frac{\partial \sigma_i}{\partial z_k} = p_i(\delta_{ik} - p_k), \quad \text{i.e.} \quad J(p) = \mathrm{Diag}(p) - pp^\top,$$

*Department of Mathematics, Centralia College, 600 Centralia College Blvd., Centralia, WA 98531, USA

and in particular $J(p)\mathbf{1} = 0$. Softmax thus descends to a diffeomorphism on the quotient $\bar{\sigma} : \mathbb{R}^n/\mathbb{R}\mathbf{1} \to \Delta^{n-1}_{>0}$ with inverse $[p] \mapsto [\log p]$.

Define $A(z) = \log \sum_i e^{z_i}$ so that $\nabla A = \sigma$ and $\nabla^2 A = J$. The Legendre dual is $A^*(p) = \sum_i p_i \log p_i = -H(p)$ and at dual pairs (z, p)

$$H(p) = A(z) - \langle p, z \rangle.$$

Equipping the quotient with the Hessian metric of $A(z) = \log \sum_i e^{z_i}$ and the simplex with the Fisher metric $\langle u, v \rangle_{F,p} = u^\top \mathrm{Diag}(1/p)\, v$, we see that $d\bar{\sigma}$ is a global Riemannian isometry. The potential A then recovers the information geometry of the softmax bundle:

$$z \xrightarrow{\nabla A} p = \sigma(z) \xrightarrow{\nabla^2 A} J(p) \equiv \text{Fisher metric.}$$

2 Replicator Dynamics and Softmax Projection

Replicator dynamics describe how populations of competing strategies evolve over time. Given a population distribution $p \in \Delta^{n-1}$ on n strategies and a payoff vector $f(p)$, the replicator equation is

$$\dot{p}_i = p_i\Big(f_i(p) - \langle f(p), p \rangle\Big), \tag{1}$$

where $f_i(p)$ is the fitness (expected payoff) of strategy i and the inner product $\langle f(p), p \rangle$ is the average fitness across the population. Strategies with above-average payoff grow in frequency, while those with below-average payoff decline. This simple law has been central in evolutionary game theory, information geometry, and population biology.

Figures (1), (2), (3), and (4) show phase portraits of four-type replicator systems on Δ^3: (1) a neutral closed orbit for the antisymmetric game, (2) an inward spiral to the uniform interior equilibrium under a negative diagonal perturbation ($\varepsilon < 0$), (3) an outward spiral to the boundary under a positive diagonal perturbation ($\varepsilon > 0$), and (4) restoration of a unique interior attractor when mutation ($\mu > 0$) is added to the outward-biased case.

Suppose logits evolve by an autonomous vector field $\dot{z} = F(z)$. By the Chain Rule, differentiating $p = \sigma(z)$ gives

$$\dot{p} = J(p)\, \dot{z} = J(p)\, F(z).$$

In components,

$$\dot{p}_i = \sum_k J_{ik}(p) F_k(z) = \sum_k p_i(\delta_{ik} - p_k) F_k(z) = p_i\Big(F_i(z) - \sum_k p_k F_k(z)\Big) = p_i\Big(F_i(z) - \langle p, F(z) \rangle\Big).$$

This is exactly the replicator equation with "fitness" $f_i(p, z) = F_i(z)$:

$$\boxed{\dot{p}_i = p_i\big(f_i - \bar{f}\big), \qquad \bar{f} = \langle p, f \rangle} \quad \Longleftrightarrow \quad \boxed{\dot{p} = J(p)\, F(z)}$$

so the induced base dynamics on the simplex are replicator dynamics.

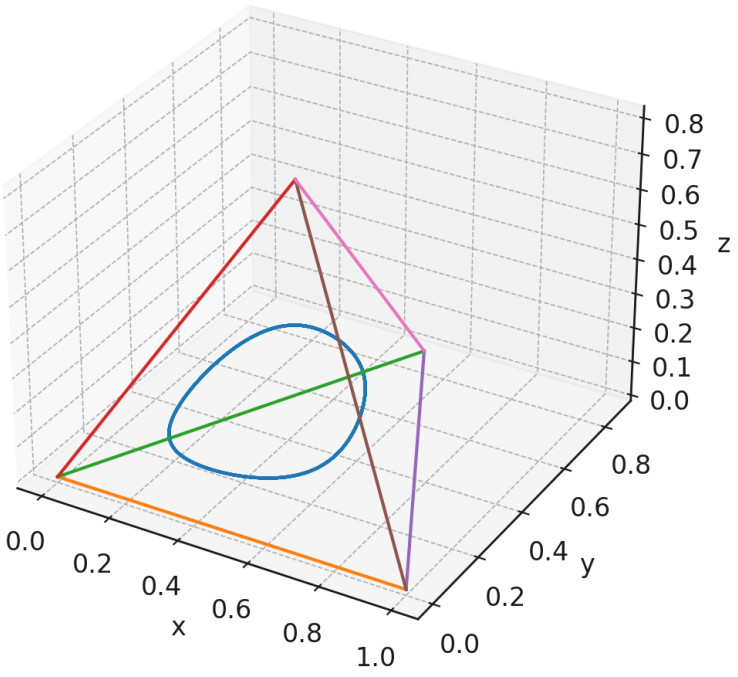

Figure 1: Four-type cyclic game, antisymmetric payoff yields a neutral closed orbit.

If F is modified by a pure gauge term $\alpha(z)\mathbf{1}$,

$$\dot{p} = J(p)\big(F(z) + \alpha\mathbf{1}\big) = J(p)F(z) + \alpha\, J(p)\mathbf{1} = J(p)F(z),$$

since $J(p)\mathbf{1} = 0$. In components this is the cancellation

$$p_i\Big((F_i + \alpha) - \langle p, F + \alpha\mathbf{1}\rangle\Big) = p_i\Big(F_i - \langle p, F\rangle\Big),$$

because $\langle p, \mathbf{1}\rangle = 1$. Any *projectable* logit field has the form

$$\dot{z} = \Phi(\sigma(z)) + \alpha(z)\mathbf{1},$$

so the base field depends only on $p = \sigma(z)$. Applying the chain rule and gauge cancellation,

$$\dot{p} = J(p)\,\Phi(p),$$

which is the canonical replicator form on the simplex. A convenient canonical lift of any simplex-tangent field $X(p)$ with $\mathbf{1}^\top X = 0$ is the Fisher/Shahshahani horizontal lift

$$\Phi_{\mathrm{hor}}(p) = \mathrm{Diag}(1/p)\, X(p), \qquad J(p)\Phi_{\mathrm{hor}}(p) = X(p).$$

3 LLM Head Mechanics as Replicator Flow

In *large language models* (LLMs), the final *head* layer computes logits from the last hidden state vector $h \in \mathbb{R}^d$:

$$z = Wh + b,$$

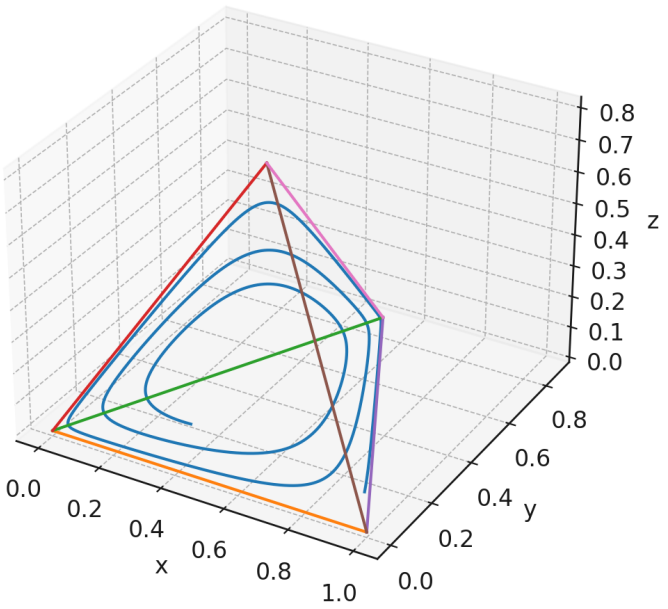

Figure 2: Diagonal perturbation $\varepsilon > 0$: outward spiral to boundary (replicator dynamics).

where $W \in \mathbb{R}^{n \times d}$ is the vocabulary projection matrix, $b \in \mathbb{R}^n$ is the bias, and n is the vocabulary size. This map is affine in h. A perturbation δh induces

$$\delta z = W \, \delta h.$$

Thus the linearization of the head is simply W, and the bias b contributes as a constant gauge shift in z-space.

If the hidden state itself evolves along some effective flow $\dot{h} = G(h)$, then the logits evolve as

$$\dot{z} = WG(h) =: F(z).$$

Hence the final head induces a vector field F on logit space, representing how logits are driven by changes in the hidden representation.

Passing through softmax, we obtain token probabilities $p = \sigma(z)$. By the general bundle construction, the induced dynamics on the simplex are

$$\dot{p} = J(p) \, F(z),$$

which is exactly a *replicator dynamic*: probability mass flows between candidate tokens proportionally to their relative advantage under F.

We may therefore summarize the head mechanics as a three-stage cycle:

1. Token $\rightarrow$ hidden state: Context tokens are embedded and processed through the transformer stack, yielding a hidden state h.

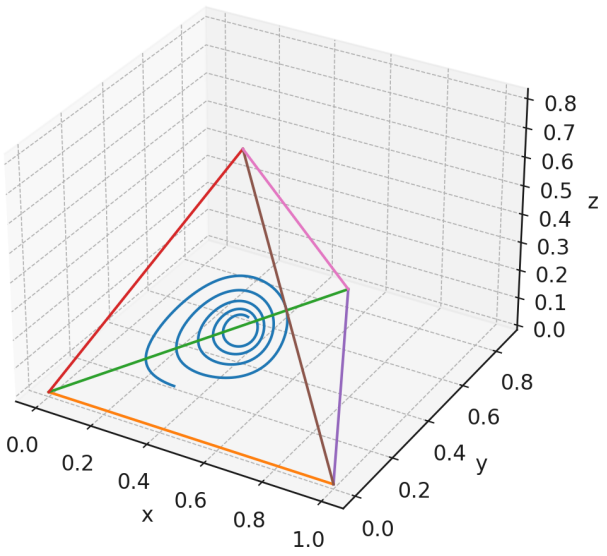

Figure 3: Diagonal perturbation $\varepsilon < 0$: inward spiral to the uniform distribution.

2. Hidden $\to$ logit flow: The affine head map $z = Wh + b$ defines a vector field $F(z)$, i.e. how candidate token scores change relative to each other.

3. Logits $\to$ probability dynamics: Softmax projects $F(z)$ down to the simplex, yielding replicator dynamics $\dot{p} = J(p)F(z)$. The trajectory $p(t)$ encodes how token probabilities evolve as competitive "populations."

The final distribution p represents the stabilized outcome of this competition. Thus, the passage from *token* to *flow* to *token* in an LLM is naturally modeled as a gauge bundle flow projecting to replicator dynamics on the probability simplex.

3.1 Geometric Intuition

The logit space can be pictured as a tangle of *spaghetti thoughts*. Each trajectory $z(t)$ in $\mathbb{R}^n$ wiggles and shifts, and different gauge choices (adding multiples of $\mathbf{1}$) correspond to different strands of spaghetti that are all physically redundant. In logit space, the geometry looks complicated: the same probabilistic story is encoded by many different lifts, see Figure 5.

When we project along the fiber direction $\mathbf{1}$—i.e. collapsing the spaghetti along the diagonal—all of these lifts fall onto a single canonical curve in the quotient space (Figure 6). This quotient is isometric with the interior of the probability simplex, where redundancy disappears.

On the simplex itself, the dynamics are governed by the replicator equation

$$\dot{p} = J(p)F(z),$$

so what looked like a messy family of spaghetti trajectories upstairs becomes a clean, intrinsic evolution of probabilities (see Figure 7). This curve is the *game* the tokens are playing:

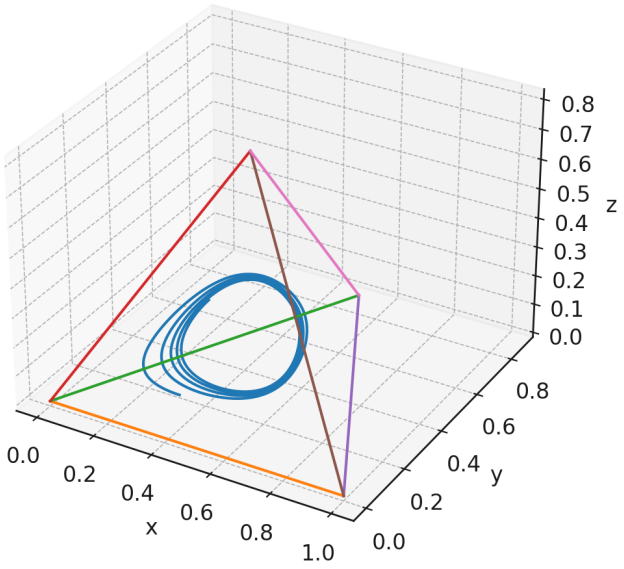

Figure 4: Outward perturbation with mutation μ: stabilized interior fixed point.

each candidate grows or shrinks relative to the others according to its advantage, and the replicator flow encodes the outcome of that competition.

4 Stacked Evolutionary Games: The Transformer as Multi-Scale Competency Architecture

So far we have analyzed the final head layer of an LLM as a softmax bundle projecting logit flows into replicator dynamics on the token simplex. But this structure is not unique to the output stage. Everywhere a softmax appears in the transformer architecture, the same geometry re-emerges, though the "game" being played changes.

Inside each attention head, a query vector q compares against keys $\{k_j\}$ to produce raw scores $z_j = \langle q, k_j \rangle$. These are normalized by a softmax,

$$p_j = \frac{e^{z_j}}{\sum_\ell e^{z_\ell}},$$

producing an attention distribution p over positions.

Thus, each attention head implements a replicator dynamic among *sequence positions*. Probability mass flows to whichever positions yield the highest advantage, concentrating or diffusing context as needed.

In mixture-of-experts or gated architectures, the gating softmax distributes weight among experts or sub-modules. Again, the geometry is identical: logits (from gating layers) live in a bundle, project to the simplex, and induce a replicator competition. Here the "tokens" are not vocabulary items but architectural pathways.

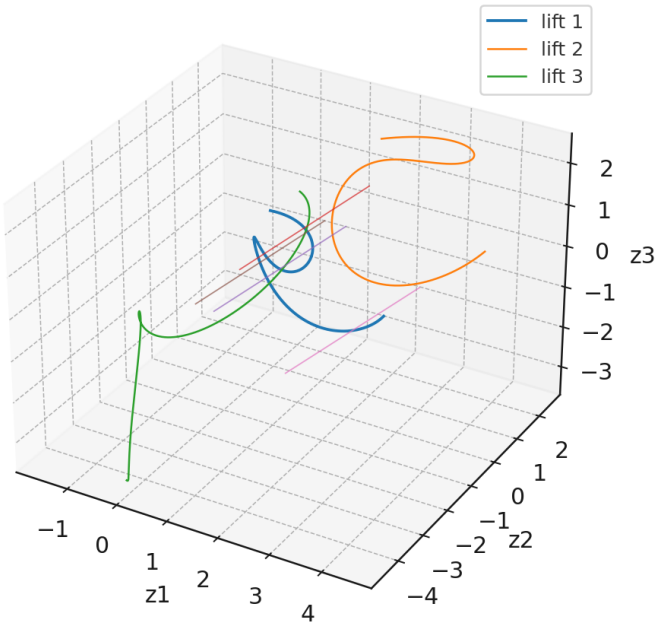

Figure 5: Logit space: three gauge-related lifts and several fiber lines.

At the last step, the hidden state h is mapped to logits $z = Wh + b$ and projected by softmax to probabilities over tokens. As shown earlier, this corresponds to a replicator game in the token simplex, the outcome of which determines the next symbol of text.

Putting this together, the transformer can be seen as a *multi-scale competency architecture* in the sense of Levin—nested, goal-directed subsystems that solve problems in their own action spaces (physiological, morphological, behavioral), exhibiting regulative plasticity across levels of organization [1, 2].

- Attention goals positions compete for influence.
- Gating goals experts or pathways compete for activation.
- Final token goal candidate symbols compete for expression.

Each softmax implements the same principal-bundle geometry—fibers, gauge, Fisher metric, replicator form—the difference is only in what is competing. By the final layer, the accumulated competitions distill context into a single probability flow over tokens.

A Temperature Effects in Softmax Dynamics

A common variant of softmax introduces a temperature parameter $\tau > 0$:

$$\sigma_i^{(\tau)}(z) = \frac{e^{z_i/\tau}}{\sum_j e^{z_j/\tau}}.$$

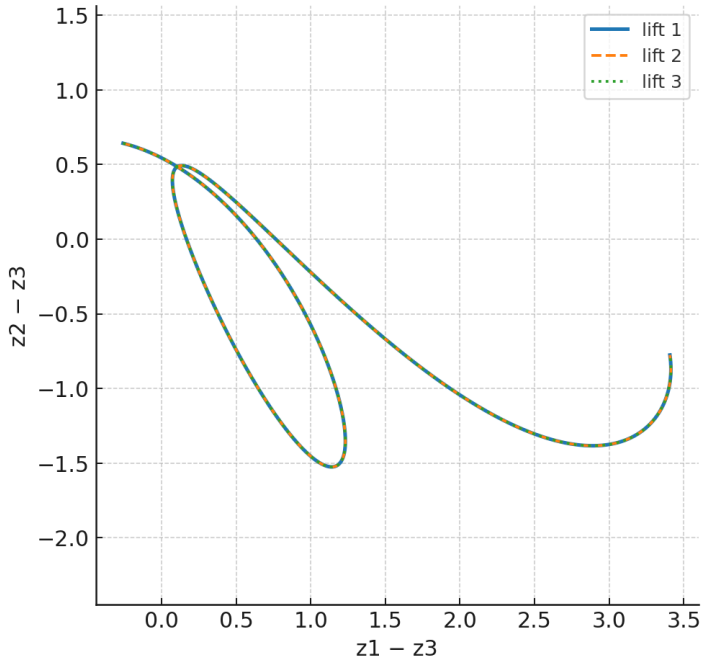

Figure 6: Gauge-free chart $(z_1 - z_3,\ z_2 - z_3)$: all lifts collapse to one curve.

The bundle structure is unchanged:

$$\sigma^{(\tau)}(z + c\mathbf{1}) = \sigma^{(\tau)}(z),$$

so the fibers are still $z + \mathbb{R}\mathbf{1}$ and the gauge group remains $(\mathbb{R}, +)$. Thus temperature modifies the geometry but does not break gauge invariance.

For $p = \sigma^{(\tau)}(z)$, the Jacobian is

$$J^{(\tau)}(p) = \frac{1}{\tau}\Big(\mathrm{Diag}(p) - pp^\top\Big).$$

Hence the Fisher–Shahshahani metric structure persists, but with an overall scale factor $1/\tau$.

Given a projectable logit field $\dot{z} = F(z)$, the induced flow on the simplex is

$$\dot{p} = J^{(\tau)}(p)F(z) = \frac{1}{\tau}J(p)F(z).$$

Thus the replicator equation is identical in form, with a rescaling of time by $1/\tau$. The temperature thus acts as an *evolutionary pressure knob*:

- $\tau \ll 1$: dynamics are sharp and fast; trajectories collapse rapidly to extreme points of the simplex (deterministic "winner-take-all").
- $\tau \gg 1$: dynamics are slow and diffuse; probability mass spreads toward the uniform distribution, dampening competition.

In this way, temperature controls the intensity of the replicator game without altering its fundamental structure.

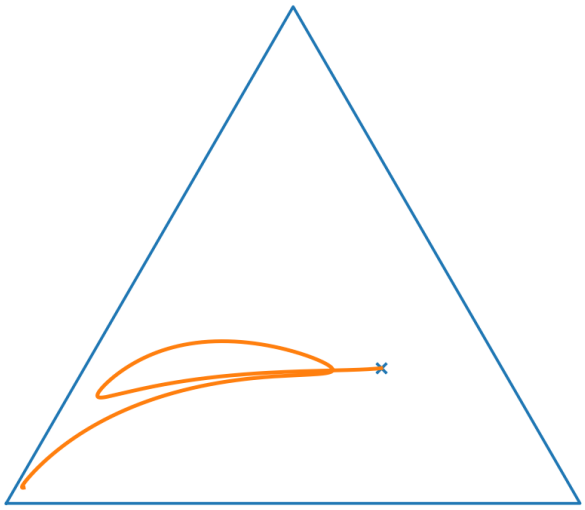

Figure 7: Base trajectory on Δ^2: common image $p(t) = \sigma(z(t))$ of all lifts.

References

[1] Levin, M. (2023). Darwin's agential materials: evolutionary implications of multiscale competency in developmental biology. *Cellular and Molecular Life Sciences*, **80**(6), 142. doi:10.1007/s00018-023-04790-z.

[2] Levin, M. (2022). Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds. *Frontiers in Systems Neuroscience*, **16**, 768201. doi:10.3389/fnsys.2022.768201.