A U-shape in dot-product attention under gradient flow

Nguyen Ngoc Khanh
khanh.nguyen.contact@gmail.com

June 5, 2026

Abstract

We prove that gradient flow on a single dot-product self-attention layer with tied unembedding, trained on cross-entropy loss in a two-token setting, exhibits a non-monotone trajectory in the Frobenius norm of its interaction matrix: starting from any *sharp-wrong* initialization in which both tokens initially attend strongly to the wrong target, there exists $T_1 \in (0, \infty)$ such that $\|M(t)\|_F^2$ is strictly decreasing on $[0, T_1]$ and diverges to $+\infty$ as $t \to \infty$. The proof rests on a row-sum conservation law for the logit matrix $L = XMX^T$, which pins $d^2 - 2$ scalar invariants of M at initialization and reduces the effective dynamics to two scalar coefficients. Borrowing the blacksmith's vocabulary, training proceeds in three phases: heating (parameter contraction on $[0, T_1]$), forging (unbounded growth for $t > T_1$), and cooling (the asymptotic regime established in prior work). We characterize the first two phases.

1 Introduction

Modern transformer training exhibits a transient phenomenon during which parameter norms briefly contract or the loss stalls before training settles into a more stable regime. Practitioners engineer around it with learning-rate warm-up schedules and norm-constrained optimizers, a rigorous account in even simplified attention models has been missing.

We prove that in the gradient flow of a single dot-product self-attention layer with tied unembedding on cross-entropy loss with two tokens, starting from a sharp-wrong initialization ($\sigma_i(0) > 1/2$ for $i = 0, 1$), there exists $T_1 \in (0, \infty)$ such that $\|M(t)\|_F^2$ is strictly decreasing on $[0, T_1]$ and diverges to $+\infty$ as $t \to \infty$ (Figure 1). To our knowledge this is the first rigorous proof of such a non-monotone trajectory in single-layer softmax attention.

2 Setup

2.1 Model

Let $d \geq 2$ and $x_0, x_1 \in \mathbb{R}^d$ be two linearly independent unit vectors with $\langle x_0, x_1 \rangle = p \in (-1, +1)$. Write $X = \begin{bmatrix} x_0, x_1 \end{bmatrix}^T \in \mathbb{R}^{2 \times d}$. A single self-attention layer with interaction matrix $M \in \mathbb{R}^{d \times d}$ is defined by

$$L = XMX^T \in \mathbb{R}^{2 \times 2}, \qquad A = \mathrm{rsm}(L) \in \mathbb{R}^{2 \times 2}, \qquad Y = AX \in \mathbb{R}^{2 \times d}$$

where rsm denotes row-wise softmax. We use tied unembedding $W = X$ (vocabulary size 2). We define the output

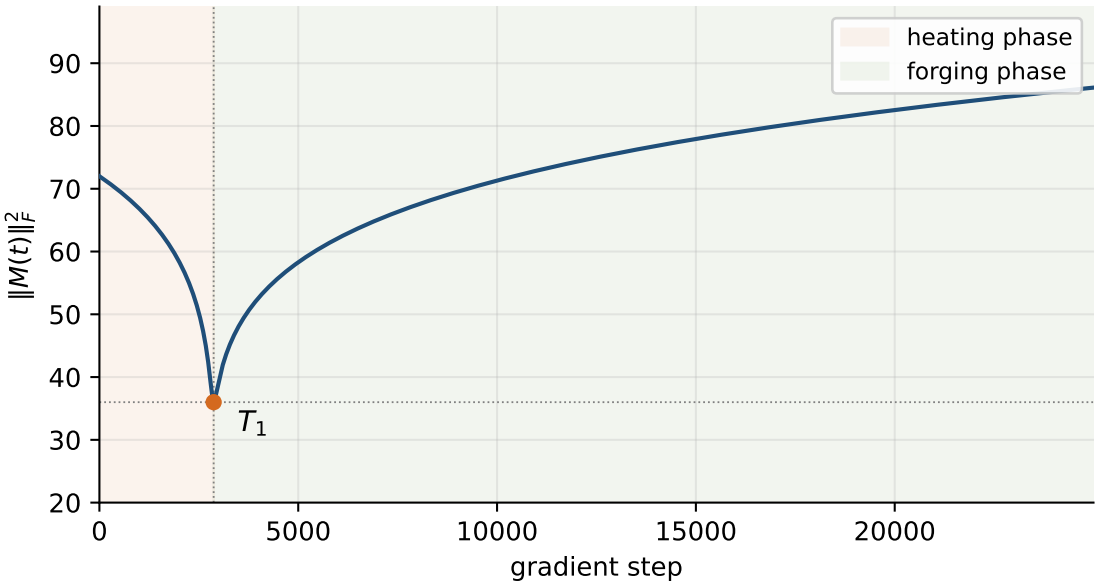

Figure 1: Squared Frobenius norm $\|M(t)\|_F^2$ along the gradient flow, for $d = 2$, $p = 0$ (orthogonal tokens), and sharp-wrong initialization $c_0 = c_1 = 6$. The trajectory strictly contracts on the heating phase $[0, T_1]$, reaches a strict trough at T_1, and diverges on the forging phase $t > T_1$.

logits $Z = YX^T \in \mathbb{R}^{2\times 2}$. Then the cross entropy loss is

$$\mathcal{L}(M) = \mathrm{ce}(Z_0, 0) + \mathrm{ce}(Z_1, 1) \tag{1}$$

where $\mathrm{ce}(z, j) = -\log \mathrm{rsm}(z)_j$ for some distribution $z \in \mathbb{R}^2$. By updating M so that it minimizes $\mathcal{L}$, we expect the attention matrix A to reach identity. In other words, we expect x_i to fully attend to itself.

2.2 Auxiliary scalars

For $i = 0, 1$, define the *wrong-direction logit gap*

$$\xi_i(M) = L_{i,1-i} - L_{i,i} \in \mathbb{R}$$

and the corresponding *wrong-direction attention*

$$\sigma_i(M) = A_{i,1-i} = \sigma(\xi_i(M)) \in \mathbb{R}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function.

2.3 Sharp-wrong initialization

At time $t = 0$, consider $M(t), \xi_i(t)$ as functions of time, we say $M(0)$ being *sharp-wrong initialized* if

$$\xi_0(0) = c_0 > 0 \quad \text{and} \quad \xi_1(0) = c_1 > 0$$

That is, $\sigma_0(0), \sigma_1(0) > 1/2$, both x_0 and x_1 attend to the wrong target.

3 U-shape

We will prove the existence of U-shape for any sharp-wrong initialization (c_0, c_1). The argument consists of four parts: closed-form loss, rank-one gradient, conservation law, and the main argument.

3.1 Loss depends on M through two scalars

Lemma 3.1 (loss function)

The loss function $\mathcal{L}$ in (1) depends on M through two scalars σ_0, σ_1

$$\mathcal{L}(\sigma_0, \sigma_1) = \log\Big(1 + e^{(1-p)(2\sigma_0 - 1)}\Big) + \log\Big(1 + e^{(1-p)(2\sigma_1 - 1)}\Big)$$

Proof. The i-th row of Y is $Y_i = (1-\sigma_i)x_i + \sigma_i x_{1-i}$. The (i,j)-th entry of Z is $Z_{i,j} = (YX^T)_{i,j} = \langle x_j, Y_i \rangle$. Therefore, the output logit gap on row i is

$$Z_{i,1-i} - Z_{i,i} = (1-p)(2\sigma_i - 1)$$

The result follows from binary cross entropy formula $\mathrm{ce}(z, j) = \log(1 + e^{z_{1-j} - z_j})$ for some distribution $z \in \mathbb{R}^2$. □

3.2 Rank-one gradient decomposition

Lemma 3.2 (rank-one structure)

The gradient $\nabla_M \mathcal{L}$ has rank at most one with the form

$$\nabla_M \mathcal{L} = 2(1-p)(\tilde{s}_0 x_0 - \tilde{s}_1 x_1) g^T$$

where $g = x_1 - x_0$ and $\tilde{s}_i = \sigma((1-p)(2\sigma_i - 1))\sigma_i(1-\sigma_i)$. In particular, $\nabla_M \mathcal{L} \in \mathcal{A} = \mathrm{span}(x_0 g^T, x_1 g^T) \subseteq \mathbb{R}^{d \times d}$

Proof. Each σ_i depends on M through ξ_i, by Lemma 3.1, $\mathcal{L}$ depends on M through ξ_0, ξ_1 with

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 2(1-p)\tilde{s}_i$$

Moreover, we can write $\xi_0(M) = x_0^T M g$ and $\xi_1(M) = -x_1^T M g$. The result follows from chain rule $\nabla_M \mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial \xi_0}\right) \nabla_M \xi_0 + \left(\frac{\partial \mathcal{L}}{\partial \xi_1}\right) \nabla_M \xi_1$. □

3.3 Conservation law of L

Theorem 3.3 (row-sum conservation)

Under gradient flow $\dot{M}(t) = -\nabla_M \mathcal{L}$ from any sharp-wrong initialization

$$(L_{i,0} + L_{i,1})(t)$$

is a constant function on $t \geq 0$.

Proof. By Lemma 3.2, for any $w \in \mathbb{R}^d$ such that $w \perp g$, we have

$$\dot{M}(t) w = -\nabla_M \mathcal{L} w = -2(1-p)(\tilde{s}_0 x_0 - \tilde{s}_1 x_1)(g^T w) = 0$$

In particular, let $w = x_0 + x_1$, then $M(t)(x_0 + x_1)$ is a constant function on $t \geq 0$ and $L(t)\mathbf{1}_2 = XM(t)(x_0 + x_1) = XM(0)(x_0 + x_1) = L(0)\mathbf{1}_2$ for $t \geq 0$ □

3.4 The (a, b) reduction and U-shape existence

By Lemma 3.2, $\dot{M}(t)$ stays in the subspace $\mathcal{A} = \mathrm{span}(x_0 g^T, x_1 g^T)$, we write

$$M(t) = M(0) + a(t)x_0 g^T + b(t)x_1 g^T \tag{2}$$

where $a(t)$ satisfies $\dot{a}(t) = -2(1-p)\tilde{s}_0(t)$ and $b(t)$ satisfies $\dot{b}(t) = +2(1-p)\tilde{s}_1(t)$ with initial conditions $a(0) = b(0) = 0$. We can also write logit gaps in term of (a, b) as follows:

$$\xi_0(t) = c_0 + \|g\|^2(a + pb) \qquad \xi_1(t) = c_1 - \|g\|^2(pa + b) \tag{3}$$

Theorem 3.4 (heating phase: U-shape existence)
Under any sharp-wrong initialization and gradient flow $\dot{M}(t) = -\nabla_M \mathcal{L}$

$$\frac{d}{dt}\|M(t)\|_F^2\Big|_{t=0} < 0$$

and there exists $T_1 \in (0, \infty)$ such that $\|M(t)\|_F$ is strictly decreasing on $[0, T_1]$ and $\|M(t)\|_F \to \infty$ as $t \to \infty$

Proof.

1. *Frobenius norm:* Using (2), expansion of $\|M(0) + ax_0 g^T + bx_1 g^T\|_F^2$ gives

$$\|M(t)\|_F^2 = \|M(0)\|_F^2 + 2(c_0 a(t) - c_1 b(t)) + 2(1-p)(a(t)^2 + 2pa(t)b(t) + b(t)^2) \tag{4}$$

   using $\langle x_0, g\rangle = p-1$, $\langle x_1, g\rangle = 1-p$ and $\langle x_0 g^T, M(0)\rangle_F = x_0^T M(0) g = +c_0$, $\langle x_1 g^T, M(0)\rangle_F = x_1^T M(0) g = -c_1$. The Hessian in (a, b) is $H = 4(1-p)\begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$ which is strictly positive definite for every $p \in (-1, 1)$. Hence, $\|M(t)\|_F^2$ is a strictly convex quadratic in (a, b).

2. *Negative initial derivative:* At $t = 0$, we have $\dot{a}(0) = -2(1-p)\tilde{s}_0(0) < 0$ and $\dot{b}(0) = +2(1-p)\tilde{s}_1(0) > 0$, differentiating (4) at $t = 0$ gives

$$\frac{d}{dt}\|M(t)\|_F^2\Big|_{t=0} = 2(c_0\dot{a}(0) - c_1\dot{b}(0)) = -4(1-p)(c_0\tilde{s}_0(0) + c_1\tilde{s}_1(0)) < 0$$

3. *Existence of the trough:* Write $f(t) = \|M(t)\|_F^2$, we will show the existence of the trough as follows:

   (a) $\|(a(t), b(t))\| \to \infty$. For finite t, $\sigma_i \in (0, 1)$, so $\tilde{s}_i > 0$, hence $\dot{a} < 0$ and $\dot{b} > 0$ strictly. By monotone convergence, $a(t) \to a_\infty \in [-\infty, 0]$ and $b(t) \to b_\infty \in [0, +\infty]$.

   Suppose the contrary that both a_∞, b_∞ are finite. Then $M(t)$ stays in some compact set $K \subseteq \mathbb{R}^{d\times d}$. On K, the function $M \mapsto \|\nabla_M \mathcal{L}(M)\|_F^2$ is continuous and positive (by Lemma 3.2, $\nabla_M \mathcal{L}$ vanishes only when $\tilde{s}_0 = \tilde{s}_1 = 0 \iff$ both $\sigma_i \in \{0, 1\} \iff$ both $\xi_i = \pm\infty$ which is excluded on the bounded set K since

each ξ_i is a linear combination of a, b). By compactness and positivity, $\delta^2 = \min_{M \in K} \|\nabla_M \mathcal{L}(M)\|_F^2 > 0$ is well-defined. Hence $\|\nabla_M \mathcal{L}(M(t))\|_F^2 \geq \delta^2$ for all $t \geq 0$, so

$$\int_0^\infty \|\nabla_M \mathcal{L}(M(t))\|_F^2 \, dt = \infty.$$

But $\frac{d}{dt}\mathcal{L} = -\|\nabla_M \mathcal{L}\|_F^2$ and $\mathcal{L} \geq 0$, so $\int_0^\infty \|\nabla_M \mathcal{L}\|_F^2 dt = \mathcal{L}(0) - \mathcal{L}_\infty < \infty$. Contradiction.

Hence at least one of a_∞, b_∞ is infinite, and $\|(a(t), b(t))\| \to \infty$.

(b) $f(t) \to \infty$. $f(t)$ as a function in (a, b) is a strictly convex quadratic. Combine with $\|(a(t), b(t))\| \to \infty$, so $f(t) \to \infty$.

(c) *Existence of a minimum $T_1 \in (0, \infty)$ such that $f(t)$ strictly decreasing on $[0, T_1]$.* Take $T_1 = \inf\{t > 0 \mid f'(t) = 0\}$. The infimum exists since $f'(0) < 0$ and $f(t) \to \infty$. The decreasing of $f(t)$ comes for free.

□

4 Discussion

The blacksmith metaphor. A useful intuition for the trajectory is the blacksmith's three-phase metaphor:

1. **Heating**: the metal (the parameter matrix M) is brought to a malleable state, the initial sharp-wrong structure is melted away and $\|M\|_F$ contracts. This is the regime $[0, T_1]$ in our theorem.

2. **Forging**: the loss signal shapes the malleable parameters, $\|M\|_F$ grows under the rank-one gradient, reorganizing itself in the direction selected by the data. This is the regime $t > T_1$ with $\|M(t)\|_F \to \infty$.

3. **Cooling**: the shape solidifies and attention sharpens onto the correct target ($\sigma_i \to 0$). This is the asymptotic regime studied in the implicit-bias literature (see related work above).

The conservation law is the load-bearing observation. Theorem 3.3 is what makes the rest of the argument work, and the proof is essentially a one-liner. Once Lemma 3.2 establishes that $\nabla_M \mathcal{L}$ has row-space $\mathrm{span}(g)$, every direction $w \perp g$ is automatically annihilated by the gradient, so $M(t)w$ is conserved for any such w. The check $\langle x_0 + x_1, g \rangle = (1 - 1) + (p - p) = 0$ exhibits $w = x_0 + x_1$ as one such direction, yielding $L(t)\mathbf{1}_2 = XM(t)(x_0 + x_1) = L(0)\mathbf{1}_2$.

The conservation admits two complementary readings:

- *Parameter side.* The gradient flow has $d^2 - 2$ scalar invariants pinned by initialization, only the projection of $M(t)$ onto the 2-D active subspace $\mathcal{A} = \mathrm{span}(x_0 g^T, x_1 g^T)$ evolves. The orthogonal component M_{frz} is frozen and contributes a fixed offset to $\|M(t)\|_F^2$ that cancels in the dip depth $f(0) - f(T_1)$.

- *Prediction side.* The row sums of the logit matrix $L = XMX^T$ never change during training. Since the softmax cross-entropy is invariant under adding a constant to a row of the logits, only the row-wise *differences* $L_{i,1-i} - L_{i,i} = \xi_i$ matter for the loss, recovering the 2-D effective dynamics in (ξ_0, ξ_1) from a coordinate-free perspective.

Why this regime is rarely studied. Most theoretical analyses of attention training assume either symmetric initialization ($\sigma_i(0) \approx 1/2$, the natural state under standard Xavier/He scaling) or initialization already in the basin of attraction

of the correct attention solution. In both cases $\|M(t)\|_F$ evolves monotonically and there is no U-shape to observe. The sharp-wrong regime sits outside both: $M(0)$ has non-trivial norm *and* is committed to the wrong attention pattern, so the gradient must first un-learn this commitment before it can build the correct one. The natural practical analogue is fine-tuning a layer whose pretraining over-committed to the wrong target: a setting most theoretical analyses skip because it requires modelling the pretraining first.

Extensions. The rank-one structure of $\nabla_M \mathcal{L}$ persists when (i) the vocabulary size $V > 2$: the additional embeddings contribute extra terms to the rate $\tilde{s}_i$ via the multi-class softmax, but the row-space of $\nabla_M \mathcal{L}$ remains $\mathrm{span}(g)$. (ii) the loss is replaced by any single-row CE / MSE / contrastive variant that factors through ξ_0, ξ_1 in the same way. In both cases the conservation law, (a, b) reduction, and U-shape existence carry over unchanged.

For $n > 2$ *input* tokens the rank-one collapse breaks, each row of A has $n - 1$ wrong-direction gaps with different row vectors, and the gradient can have rank up to $n(n-1)/2$. Whether a U-shape persists per row in this richer regime is open.

A Toy example: $X = I_2$

The simplest specialization is $X = I_2$ (so $p = 0$), giving $L = M$ and $A = \mathrm{rsm}(M)$. The cross-entropy on attention rows decouples and one can prove the U-shape directly, without invoking the active-subspace machinery.

With $\mathcal{L} = -\log A_{00} - \log A_{11}$, the row-decoupling is exact: $\mathcal{L} = \mathcal{L}_0(M_{00}, M_{01}) + \mathcal{L}_1(M_{10}, M_{11})$. Analyzing row 0 with $u := M_{00}$, $v := M_{01}$ and initialization $M(0) = \begin{bmatrix} 0 & c \\ c & 0 \end{bmatrix}$:

- $\mathcal{L}_0 = \log(1 + e^{v-u})$, so $\dot{u} = \sigma(v - u)$, $\dot{v} = -\sigma(v - u)$.
- $\dot{u} + \dot{v} = 0$ and $u(0) + v(0) = 0 + c = c$, so $u + v = c$ is conserved (a special case of Theorem 3.3).
- Substituting $v = c - u$: $\rho^2 := u^2 + v^2 = u^2 + (c - u)^2 = 2(u - c/2)^2 + c^2/2$, a strictly convex parabola in u with minimum $\rho = c/\sqrt{2}$ at $u = c/2$ and equal endpoint values $\rho = c$ at $u \in \{0, c\}$.
- Since $\sigma > 0$, $\dot{u} > 0$ and u traverses $[0, c]$ from 0 to c, by the intermediate value theorem there are unique times T_1, T_2 with $u(T_1) = c/2$ and $u(T_2) = c$, and ρ is strictly decreasing on $[0, T_1]$ and strictly increasing on $[T_1, T_2]$.

Row 1 gives the same trajectory by symmetry, so $\|M(t)\|_F^2 = 2\rho(t)^2$ and

$$\|M\|_F \;:\; c\sqrt{2} \;\longrightarrow\; c \;\longrightarrow\; c\sqrt{2},$$

matching the predicted trough $\|M\|_F = c/(1 + p) = c$ at $p = 0$.