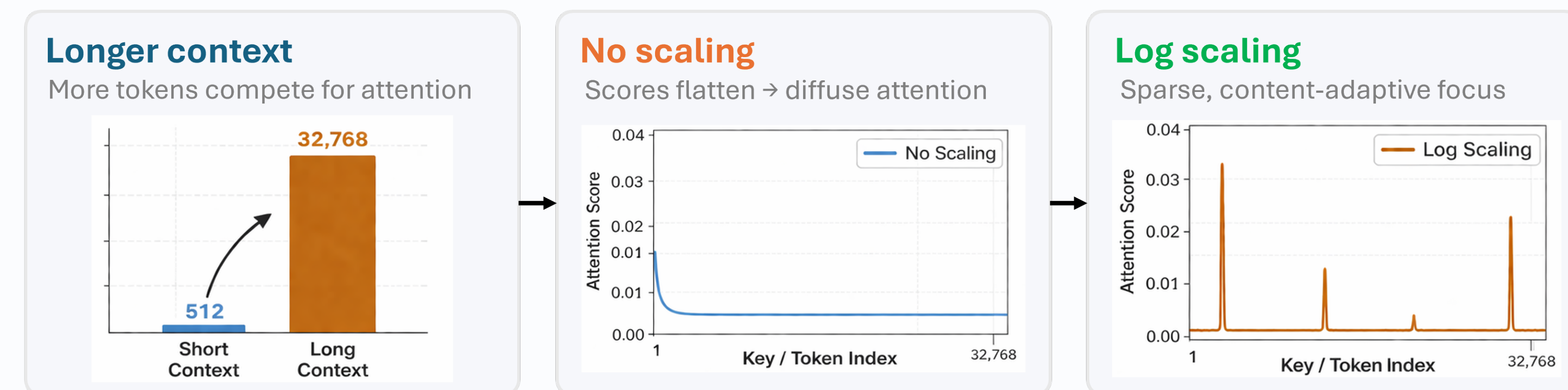


Why long context breaks attention

Main takeaway

- Too little scaling makes attention uniform and causes rank-collapse;
- Too much scaling makes attention identity and ignores interactions;
- The critical order is $\beta_n = \Theta(\log n)$, which preserves sparse, content-adaptive attention in long contexts.

As the context length n grows, attention scores flatten toward uniformity, so each token spreads mass across all neighbors instead of focusing selectively. The result is **rank-collapse**: token representations contract toward each other.



Transformer model

$$y_i = N(x_i) = \frac{x_i}{\|x_i\|} \in \mathbb{S}^{d-1}, \quad a_{ij} = \beta_n \langle y_i, y_j \rangle, \quad Z_i = \sum_{k=1}^n e^{a_{ik}}, \quad A_{ij} = \frac{e^{a_{ij}}}{Z_i}.$$

We update tokens by the formula

$$x'_i = \text{ATT}(y_i) + \alpha x_i, \quad \text{with} \quad \text{ATT}(y_i) = \sum_{j=1}^n A_{ij} y_j.$$

Residuals soften collapse, but they do *not* remove the phase transition. The key question is the right order of β_n as $n \rightarrow \infty$ so that tokens exhibit different collapsing phases.

Main theorem, forward pass: $\beta_n \asymp \log n$

Simplex case

If normalized tokens satisfy $\langle y_i, y_j \rangle = \rho$ for all $i \neq j$, and $\beta_n = \gamma \log n$, then the sharp threshold is

$$\gamma^* = \frac{1}{1-\rho}.$$

Denote $y'_i = N(x'_i)$. Without residual connections ($\alpha = 0$), as $n \rightarrow +\infty$,

$$\langle y'_i, y'_j \rangle \rightarrow \begin{cases} 1, & \gamma < \gamma^*, \text{ tokens collapse to one point in one step,} \\ \frac{4\rho}{1+3\rho}, & \gamma = \gamma^*, \text{ tokens collapse but not to a single point,} \\ \rho, & \gamma > \gamma^*, \text{ tokens do not move.} \end{cases}$$

With residual connections ($\alpha > 0$) and $\|x_i\| = 1$,

$$\langle y'_i, y'_j \rangle \rightarrow \begin{cases} \frac{\rho(\alpha+1)^2}{\alpha^2 + 2\alpha\rho + \rho}, & \gamma < \gamma^*, \\ \frac{\rho(\alpha+1)^2}{\alpha^2 + \alpha(1+\rho) + \frac{1+3\rho}{4}}, & \gamma = \gamma^*, \\ \rho, & \gamma > \gamma^*. \end{cases}$$

So attention moves from **strong contraction** to a **critical sparse regime** and then to an **identity-like regime**, depending on γ and the threshold γ^* .

Beyond the simplex

When token similarities vary in $[\rho_1, \rho_2]$, the same order persists: attention is contractive below $\frac{1}{1-\rho_1}$, identity-like above $\frac{1}{1-\rho_2}$, and can exhibit a middle sparse phase in between.

Main theorem, backward pass

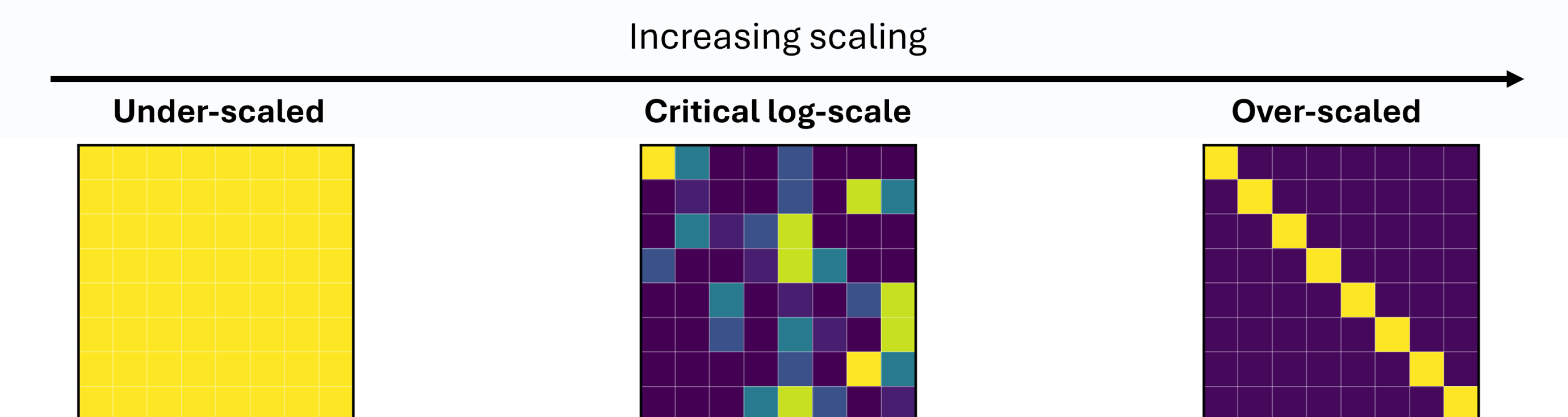
Backward pass

Let $X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^{nd}$. The normalized Jacobian norm undergoes the same phase transition. Without residual connections ($\alpha = 0$) and $\|x_i\| = 1$, as $n \rightarrow +\infty$,

$$\frac{1}{nd} \|\nabla_X X'\|_F^2 \rightarrow \begin{cases} 0, & \gamma < \gamma^*, \text{ Jacobian vanishes,} \\ \frac{1}{4} \left(1 - \frac{1}{d}\right), & \gamma = \gamma^*, \\ \left(1 - \frac{1}{d}\right), & \gamma > \gamma^*, \text{ Jacobian is nearly identity.} \end{cases}$$

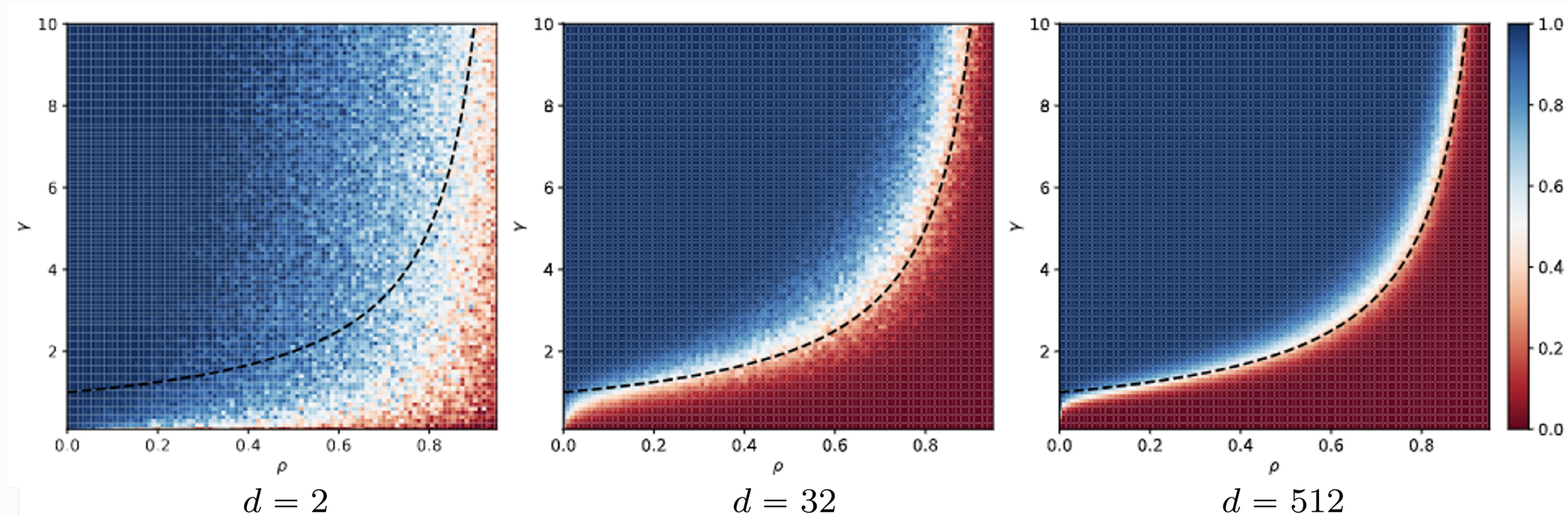
Practical reading. The forward and backward plots tell the same story: the same scaling that keeps token geometry informative also keeps gradients trainable.

Scaling regimes and practical implications



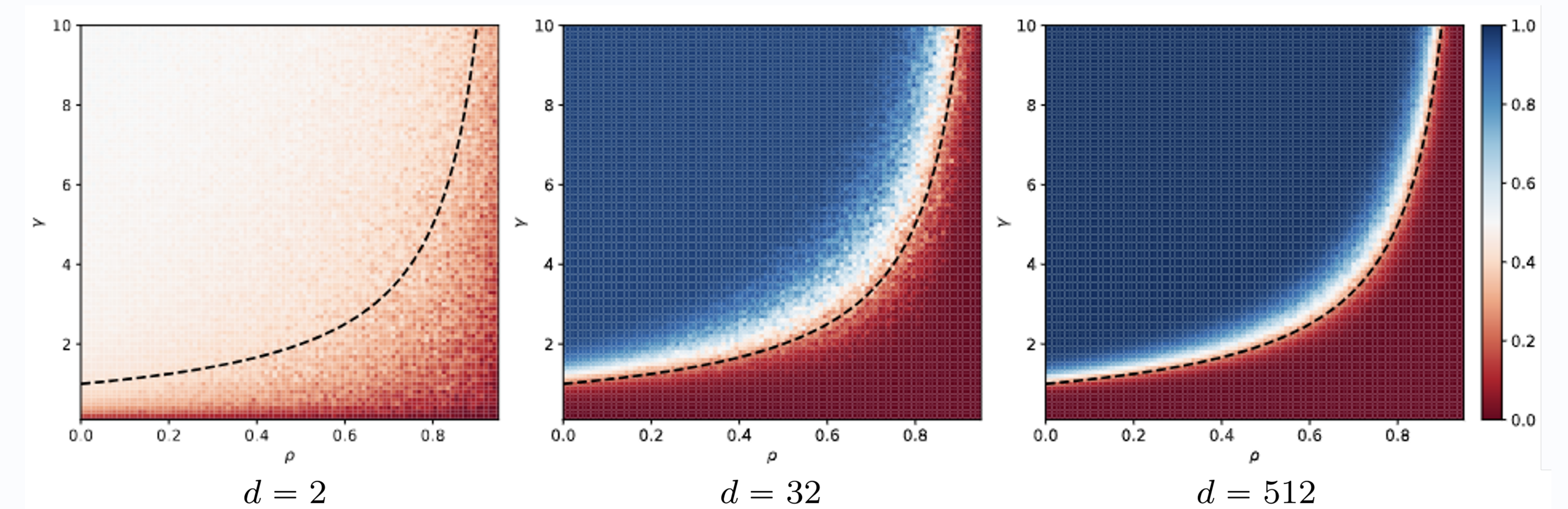
- **Subcritical:** diffuse attention, rank-collapse, vanishing gradients.
- **Critical:** content-adaptive attention with meaningful interactions.
- **Supercritical:** near-identity attention and weak mixing.
- **Logarithmic attention scaling models:** Qwen, YaRN.
- The exact constant depends on token geometry: as token similarities concentrate, the transition becomes sharper.

Figure 1: forward dynamics



Heatmaps of the input-to-output angle ratio λ . The dashed boundary is $\gamma = \frac{1}{1-\rho}$. As d grows, the transition sharpens and aligns with the theoretical phase boundary.

Figure 2: backward dynamics



Heatmaps of the normalized gradient norm η over the same (ρ, γ) grid. Small γ drives the gradient norm close to zero; larger γ preserves signal. The same boundary that controls token contraction also controls trainability.