"

A PHASE TRANSITION OF ATTENTION COLLAPSE

ANIKET DESHPANDE

Abstract

In this brief note, we analyze a toy model of scaled dot-product attention in which the logits are conditionally i.i.d. Gaussian, and the softmax inverse-temperature β controls sharpness. By matching a bulk log-sum-exp approximation for $\log Z_\beta$ with the extreme-value scaling of the maximum logit M_N, we obtain a critical inverse temperature $\beta_c(N,\sigma) = \sigma^{-1}\sqrt{2\log N}$. Below β_c, the maximum attention weight w_{max} vanishes as $N \to \infty$, resulting in a *diffuse* phase. Above β_c, attention *condenses* onto $O(1)$ keys with $w_{max} = \Theta(1)$, approaching $w_{max} \to 1$ only in the deeper low-temperature limit $\beta/\beta_c \to \infty$. We interpret the Transformer's $d^{-1/2}$ scaling as preventing trivial collapse at fixed β and connect the phenomenon to classic condensation in random-energy models.

Let us define a toy attention model. Let N be the sequence length (the number of competing keys). The Transformer head dimension d is the embedding and key dimension. Lastly, we define inverse temperature $\beta \geq 0$, a softmax sharpness. Let the query be a (possibly random) vector $q \in \mathbb{R}^d$. Let the keys be $k_1, \ldots, k_N \in \mathbb{R}^d$. Using this, let us define scaled dot-product logits:

$$U_j := \frac{1}{\sqrt{d}} q^T k_j, \quad j = 1, \ldots, N, \tag{1}$$

and define softmax attention weights

$$w_j := \frac{\exp(\beta U_j)}{\sum_{\ell=1}^N \exp(\beta U_\ell)}, \quad \sum_{j=1}^N w_j = 1, \, w_j \geq 0. \tag{2}$$

With this, let us define the partition function

$$Z_\beta := \sum_\ell^N \exp(\beta U_\ell). \tag{3}$$

We wish to examine *attention collapse*. A clean order parameter is the maximum attention weight

$$w_{max} := \max_{j \in [N]} w_j. \tag{4}$$

A *diffuse* or uniform attention would result in the maximum attention weight vanishing as $N \to \infty$. Typically, $w_{max} \sim 1/N$. A *collapsed* attention would mean $w_{max} \to 1$; almost all mass is on one key. We will show a sharp threshold in β separating these regimes. Assume the keys are i.i.d. standard Gaussian $k_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and q is independent of $\{k_j\}$. First, let us condition on q. Because k_j is normally distributed,

$$q^T k_j \mid q \sim \mathcal{N}(0, \|q\|^2) \implies U_j \mid q \sim \mathcal{N}(0, \sigma^2),$$

with the variance defined as $\sigma^2 := \|q\|^2/d$. Moreover, conditional on q, the U_j are i.i.d. Let $M_N := \max_{j \in [N]} U_j$. Then, always,

$$\exp(\beta M_N) \leq Z_\beta \leq N \exp(\beta M_N). \tag{5}$$

Taking logarithms, we arrive at

$$\beta M_N \leq \log Z_\beta \leq \beta M_N + \log N.$$

This is simply because the largest term is at most the sum, and the sum is at most N times the largest term. The phase transition emerges from a competition. In the *bulk regime*, many terms contribute to Z_β. In the *maximum regime*, one extreme term dominates Z_β. Let us compute the expectation of the partition function using the moment

Now, we solve for the critical inverse temperature β_c by matching bulk and maximum approximations. The transition occurs when the bulk and max approximations are of the same order:

$$\log N + \frac{1}{2}\beta^2\sigma^2 \approx \beta\sigma\sqrt{2\log N}. \tag{7}$$

We rearrange and find a condition for when the expression vanishes

$$0 \approx \frac{1}{2}\beta^2\sigma^2 - \beta\sigma\sqrt{2\log N} + \log N = \frac{1}{2}\sigma^2\left(\beta - \frac{1}{\sigma}\sqrt{2\log N}\right)^2.$$

This vanishes at exactly

$$\beta_c(N,\sigma) = \frac{1}{\sigma}\sqrt{2\log N}. \tag{8}$$

This is the *critical* inverse temperature for softmax condensation over N Gaussian logits, analogous to the freezing transition in random energy models [2]. Now, let us show that the order parameter jumps from 0 to 1. Pick $j^* := \arg\max_j U_j$, so that $U_{j^*} = M_N$. Then,

$$w_{max} = w_{j^*} = \frac{\exp\left(\beta M_N\right)}{Z_\beta}. \tag{9}$$

When $w_{max} \to 0$, we are below criticality. We use the bulk approximations $Z_\beta \approx N\exp^{(\beta^2\sigma^2)/2}$ and $M_N \approx \sigma\sqrt{2\log N}$:

$$w_{max} \approx \exp\left(\beta\sigma\sqrt{2\log N} - \log N - \frac{1}{2}\beta^2\sigma^2\right). \tag{10}$$

Let us define the term inside the exponential as $\Phi(\beta)$ and complete the square,

$$\Phi(\beta) = -\frac{1}{2}\sigma^2\left(\beta - \frac{1}{\sigma}\sqrt{2\log N}\right)^2 = -\frac{1}{2}\sigma^2\left(\beta - \beta_c\right)^2 \leq 0.$$

If $\beta = (1-\delta)\beta_c(N)$ with fixed $\beta > 0$, then $\Phi(\beta) \sim -\delta^2\log N$ and $w_{max} \to 0$ as $N \to \infty$. More precisely, the maximum weight stays vanishingly small compared to 1 and attention remains diffuse. Now, let us consider the regime above criticality. When $\beta > \beta_c$, the partition function is no longer controlled by the bulk of $O(N)$ typical logits, but instead by the extreme tail. So, only the largest few U_j contribute appreciably to

$$Z_\beta = \sum_{j=1}^{N}\exp(\beta U_j).$$

Although the top logit $M_N = \max_j U_j$ is separated from the bulk by a gap of order $\sqrt{\log N}$, the gaps between the top few logits are much smaller, so we should not generally expect a deterministic single-winner limit $w_{max} \to 1$ and fixed $\beta/\beta_c > 1$. Instead, writing $j^* = \arg\max_j U_j$, we have the exact identity:

$$w_{max} := w_{j^*} = \frac{\exp(\beta M_N)}{\sum_{j=1}^{N}\exp(\beta U_j)} = \frac{1}{1 + \sum_{j\neq j^*}\exp(-\beta(M_N - U_j))},$$

and for $\beta > \beta_c$, the sum receives non-negligible contributions only from a finite number of near-maximum logits. Hence w_{max} becomes $\Theta(1)$ (bounded away from 0) and attention concentrates on the

top $O(1)$ keys. This is the collapsed or *condensed* phase. In the deeper low-temperature limit $\beta/\beta_c \to \infty$, the softmax approaches a hard argmax and $w_{max} \to 1$.

$$w_{\max} = \frac{1}{1 + \sum_{j \neq j^*} \exp\left(-\beta(M_N - U_j)\right)} = \Theta(1), \qquad \beta > \beta_c. \tag{11}$$

Thus, the model exhibits an attention collapse transition at the critical inverse temperature β_c. Let us substitute in Transformer variables and explore dependence on N and d. Recall that $\sigma^2 = \|q\|^2/d$. If q is typical isotropic with $\|q\|^2 \approx d$, then $\sigma \approx 1$ and

$$\beta_c(N) \approx \sqrt{2\log N} \tag{12}$$

for a single attention head with Gaussian-like logits. If we *remove* the Transformer scaling and instead use logits $U_j = q^T k_j$, then $\sigma^2 \approx \|q\|^2 \approx d$, so

$$\beta_c \approx \sqrt{\frac{2\log N}{d}}.$$

For any fixed $\beta = O(1)$, large d would push far above β_c, resulting in *trivial collapse*. This is why the $d^{-1/2}$ factor is essential. Finally, in more realistic long-context attention-layer models (with LayerNorm and residual structure and evolving tokens), the critical scaling can shift. For example, the critical scaling can shift as $\beta_n \asymp \log n$ in the tractable model studied by [1]. However, the mechanism is the same: a sharp regime change governed by how β compares to the extreme value scale of N competing logits. If one key has a deterministic advantage m[1] while the others are $U_j \sim \mathcal{N}(0, \sigma^2)$, then the target weight is

$$w_* = \frac{\exp\left(\beta m\right)}{\exp\left(\beta m\right) + \sum_{j=2}^{N} \exp\left(\beta U_j\right)}. \tag{13}$$

In the collapsed regime, $\sum \exp(\beta U_j) \approx \exp(\beta M_N)$, so

$$w_* \approx \frac{1}{1 + \exp\left(\beta(M_N - m)\right)}.$$

Successful retrieval occurs when the signal advantage m is strong enough to outcompete the noise floor. Specifically, the condition is

$$m > \sigma\sqrt{2\log N}.$$

When this inequality fails, we have *condensation on noise*: the softmax collapses onto a random key from the noise distribution, resulting in hallucination. When the inequality holds, we have *condensation on signal*: the target key with advantage m captures most of the attention mass, enabling successful retrieval. Alternatively, increasing β can sharpen the separation even when the signal advantage is marginal.

---

[1] the target logit $= m$.

References

[1] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. *Critical attention scaling in long-context transformers.* arXiv:2510.05554, 2025. `https://arxiv.org/abs/2510.05554`

[2] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation.* Oxford University Press, 2009. (See Chapter 5: The Random Energy Model.) ISBN: 978-0198570837. `https://global.oup.com/academic/product/information-physics-and-computation-9780198570837`

[3] Andrew B. Nobel. *Gaussian Extreme Values.* Lecture notes, March 2023. `https://nobel.web.unc.edu/wp-content/uploads/sites/13591/2024/04/Gaussian_Extremes.pdf`