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 $\mathsf{softmax}$ inverse-temperature β controls sharpness. By matching a bulk (high-temperature) log-sum-exp approximation for the *quenched* log-partition $\log Z_\beta$ 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 keeping logits at $O(1)$ variance (preventing trivial collapse at fixed β and large d) while also stabilizing gradients, and connect the phenomenon to classic freezing/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 $\mathsf{softmax}$ 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 $\mathsf{softmax}$ attention weights

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

With this, let us define the partition function

$$Z_\beta := \sum_{\ell=1}^N \exp\left(\beta U_\ell\right). \tag{3}$$

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* attention would result in the maximum attention weight vanishing as $N \to \infty$ (typically $w_{max} \sim 1/N$). A *collapsed* attention means $w_{max} = \Theta(1)$, with $w_{max} \to 1$ only in a deep low-temperature limit where $\mathsf{softmax}$ approaches a hard argmax. We will show a sharp threshold in β separating these regimes.

Date: January 5, 2026.

These notes supplement a blog post on `aniketdeshpande.com`.

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 conditional 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,

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

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*, a finite number of extreme terms dominate Z_β.
It is useful to separate *annealed* and *quenched* log-partitions. Define

$$A_\beta := \log \mathbb{E}\left[Z_\beta \mid q\right], \qquad Q_\beta := \mathbb{E}\left[\log Z_\beta \mid q\right].$$

By Jensen, $Q_\beta \leq A_\beta$. We compute the annealed partition function using the moment

$$\mathbb{E}[Z_\beta \mid q] = N \mathbb{E}[\exp\left(\beta U\right)] = N \exp\left(\frac{1}{2}\beta^2\sigma^2\right), \quad U \sim \mathcal{N}(0, \sigma^2),$$

so

$$A_\beta = \log N + \frac{1}{2}\beta^2\sigma^2. \tag{6}$$

To relate this to the *typical* (quenched) behavior of $\log Z_\beta$, note that the map

$$(u_1, \ldots, u_N) \mapsto \log\Big(\sum_{j=1}^{N} e^{\beta u_j}\Big)$$

is β-Lipschitz in the Euclidean norm: its gradient is $\beta(w_1, \ldots, w_N)$, so $\|\nabla \log Z_\beta\|_2 = \beta\|w\|_2 \leq \beta$. Hence, conditional on q, standard Gaussian concentration for Lipschitz functions implies that $\log Z_\beta$ fluctuates around Q_β by at most $O(\beta\sigma)$ with overwhelming probability [5]. In the regime where $\log Z_\beta$ is order $\log N$, these fluctuations are lower-order. In particular, in the high-temperature/bulk regime (below the critical point defined below), the REM computation shows that $Q_\beta = A_\beta + o(\log N)$, i.e. annealed and quenched free energies match at leading order [3]. Thus, in the bulk regime, A_β is a correct *log-scale* approximation for $\log Z_\beta$:

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

For N i.i.d. Gaussians, the maximum satisfies the classic scale [4]

$$M_N \approx \sigma\sqrt{2\log N}$$

up to lower-order corrections (e.g., of order $\sigma/\sqrt{\log N}$). If the extreme tail dominates, then

$$\log Z_\beta \approx \beta M_N \approx \beta\sigma\sqrt{2\log N}.$$

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:

$$\log N + \frac{1}{2}\beta^2\sigma^2 \approx \beta\sigma\sqrt{2\log N}. \tag{8}$$

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

$$\beta_c(N,\sigma) = \frac{1}{\sigma}\sqrt{2\log N}. \tag{9}$$

This is the *critical* inverse temperature for $\mathsf{softmax}$ condensation over N Gaussian logits, analogous to the freezing transition in random energy models [3]. (More formally, in the low-temperature phase the ordered Gibbs weights converge to a Poisson–Dirichlet random mass partition in the REM class [6].)

Now, let us show that the order parameter changes from vanishing to $\Theta(1)$. Pick $j^* := \arg\max_j U_j$, so that $U_{j^*} = M_N$. Then,

$$w_{max} = w_{j^*} = \frac{\exp\left(\beta M_N\right)}{Z_\beta}. \tag{10}$$

When $w_{max} \to 0$, we are below criticality. Using the bulk approximation $\log Z_\beta \approx \log N + \frac{1}{2}\beta^2\sigma^2$ and $M_N \approx \sigma\sqrt{2\log N}$ gives

$$w_{max} \approx \exp\left(\beta\sigma\sqrt{2\log N} - \log N - \frac{1}{2}\beta^2\sigma^2\right). \tag{11}$$

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

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

A clean way to interpret the asymptotics is to compare β to β_c. Fix $t \in (0,1)$ and set $\beta = t\,\beta_c(N,\sigma)$. Then $\Phi(\beta) = -(1-t)^2\log N$ and $w_{max} \approx N^{-(1-t)^2} \to 0$ as $N \to \infty$. In particular, for any fixed $\beta = O(1)$ and $N \to \infty$, we have $\beta/\beta_c \to 0$ 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 near-maximum spacings are much smaller (so we should not generally expect a deterministic single-winner limit at fixed $\beta/\beta_c > 1$) [4]. Writing $j^* = \arg\max_j U_j$, we have the exact identity

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

For $\beta > \beta_c$, the sum receives non-negligible contributions only from a finite number of near-maximum logits, and one enters the condensed phase where w_{max} is order one. More sharply, in the REM universality class the full vector of decreasing weights has a non-degenerate Poisson–Dirichlet limit, so w_{max} converges in law to a random variable in $(0,1)$ (hence $\Theta(1)$, but not generically $\to 1$) [6]. In the deeper low-temperature limit $\beta/\beta_c \to \infty$, $\mathsf{softmax}$ approaches a hard argmax and $w_{max} \to 1$.

$$w_{\max} = \frac{1}{1 + \sum_{j \neq j^*} \exp\left(-\beta(M_N - U_j)\right)} = \Theta(1), \qquad \beta > \beta_c. \tag{12}$$

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 $\|q\|^2 \approx d$ (e.g., by norm concentration for large d), then $\sigma \approx 1$ and

$$\beta_c(N) \approx \sqrt{2\log N} \tag{13}$$

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* driven by noise extremes. This is one reason the $d^{-1/2}$ factor is essential: it keeps logit variance $O(1)$ across head sizes. In addition, the original Transformer motivation emphasizes gradient stability: without scaling, dot products grow in magnitude with d_k, pushing $\mathsf{softmax}$ into saturation and producing very small gradients [1]. 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 [2]. 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

$$w_* = \frac{\exp(\beta m)}{\exp(\beta m) + \sum_{j=2}^{N} \exp(\beta U_j)}. \tag{14}$$

Successful retrieval occurs when the signal advantage outcompetes the noise floor, but the sharp condition depends on the phase.
In the diffuse (bulk) regime, $\log \sum_{j=2}^{N} e^{\beta U_j} \approx \log N + \frac{1}{2}\beta^2\sigma^2$, so

$$w_* \approx \exp\left(\beta m - \log N - \frac{1}{2}\beta^2\sigma^2\right).$$

Thus w_* remains non-negligible only if

$$m \gtrsim \frac{\log N}{\beta} + \frac{1}{2}\beta\sigma^2.$$

[1] the target logit $= m$.

In the condensed (extreme) regime, $\sum_{j=2}^{N} \exp(\beta U_j)$ is dominated by a finite number of near-maximal noise logits, and the relevant comparison is to M_N:

$$w_* \approx \frac{1}{1 + \exp\big(\beta(M_N - m)\big) \cdot \Theta(1)}.$$

Thus, retrieval requires m to exceed the top noise level by at least an $O(1/\beta)$ margin. In the hard-argmax limit $\beta/\beta_c \to \infty$, this reduces to the extreme-value inequality

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

i.e. the signal must beat the noise maximum. When this inequality fails, we have *condensation on noise*: increasing β sharpens the argmax *toward* the largest noise key, producing a “hallucination” winner. When it holds, we have *condensation on signal*: the target key captures most of the attention mass.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need.* NeurIPS, 2017. https://arxiv.org/abs/1706.03762

[2] 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

[3] 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

[4] 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

[5] Marek Biskup. *Lecture 6: Concentration for the maximum.* PIMS notes, June 2017. https://www.math.ucla.edu/~biskup/PIMS/PDFs/lecture6.pdf

[6] Gérard Ben Arous, Véronique Gayrard, and A. Kuptsov. *A New REM Conjecture.* Preprint, May 2007. https://math.nyu.edu/faculty/benarous/Publications/benarous_72.pdf