Is Higher Singular Value Entropy of Matrix Parameters Always Better?

Su Jianlin

2026-05-29

1 Introduction

In last year's technical report "Muon is Scalable for LLM Training", to compare differences between models trained by Muon and Adam, we introduced the concept of "singular value entropy". We observed that the matrix parameters trained by Muon generally have higher singular value entropy than those trained by Adam, and we used this as empirical evidence supporting "Muon makes better use of model parameters".

However, is higher singular value entropy always better? Is there an optimal singular value entropy? Or conversely, if we must pre-specify the singular value entropy of matrix parameters, what value should we choose?

2 Concept Review

First, we review some concepts. We know that singular values of a matrix are non-negative, so we can normalize them in some way, such as direct normalization or squared normalization:

$$p_i = \frac{\sigma_i}{\sum_{j=1}^n \sigma_j} \qquad \text{or} \qquad p_i = \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2} \tag{1}$$

This yields an n-dimensional probability distribution $\boldsymbol{p} = (p_1, p_2, \cdots, p_n)$. Given such a distribution, we can compute related variables, such as entropy:

$$H(\boldsymbol{p}) = -\sum_{i=1}^n p_i \log p_i \tag{2}$$

This is the notion of "singular value entropy". We know that maximum entropy corresponds to the uniform distribution, so singular value entropy measures how uniformly distributed the singular values are. However, entropy is not unique; the logarithmic expectation form above is known as "Shannon entropy". There are other types, such as Rényi entropy:

$$H_q(\boldsymbol{p}) = -\frac{1}{q-1}\log \sum_{i=1}^n p_i^q \tag{3}$$

Shannon entropy is a special case corresponding to $q \to 1$. For other q, the entropy also achieves maximum $\log n$ under uniform distribution and minimum 0 under one-hot distribution. Singular value entropy is essentially equivalent to the matrix's effective rank, and can also be interpreted as measuring the sparsity of singular values (see "How to Measure Data Sparsity?").

3 Problem Reformulation

Now, how can we turn the question "Is higher singular value entropy always better?" into a computable proposition to obtain a reference answer?

Intuitively, since current models are generally over-parameterized, there exist many redundant degrees of freedom, which allows us to impose constraints on model parameterssuch as constraining their norm or singular value entropywithout significantly degrading performance. After imposing such constraints, the degrees of freedom are naturally reduced. We can empirically assume that the remaining degrees of freedom correlate with model capacity, which forms the basis for searching for an "optimal singular value entropy".

Thus, given a fixed singular value entropy, how many degrees of freedom remain in the matrix? Consider two extreme cases: The first is minimal entropy, where only one singular value is non-zero, making the matrix rank-1, with degrees of freedom not exceeding $2n$. The second is maximal entropy, where all non-zero singular values are equal, resulting in a scalar multiple of an orthogonal matrix, with degrees of freedom approximately $n^2/2$. Thus, maximal entropy appears better than minimal entropy, but could an intermediate value be even better?

We can also abstract away from the matrix context and directly examine the "entropy density" of singular value distributions. The maximum entropy of an n-dimensional distribution is $\log n$, achievable only by the uniform distribution; the minimum is 0, corresponding to one-hot distributions, with n possible choices. Clearly, the "density" at these boundaries is lowthey represent only a small subset of all distributionssuggesting that the optimal value likely lies in an intermediate region.

This reformulation transforms our question into: If we uniformly sample n-dimensional probability distributions and compute their entropies, what does the entropy distribution look like? Where is its peak density? The point of highest probability density corresponds to the neighborhood containing the most distributions, which we consider to represent the highest expressive power.

4 Geometric Perspective

Some readers may find "uniformly sampling n-dimensional distributions" unintuitivetypically we sample from a given distribution, not the distribution itself. However, by temporarily suspending $\boldsymbol{p}$'s probabilistic interpretation, we gain a clear geometric picture.

An n-dimensional probability distribution $\boldsymbol{p}$ is an n-dimensional vector satisfying:

$$0 \leq p_i \leq 1 \quad (\forall i = 1, 2, \cdots, n), \qquad \sum_{i=1}^{n} p_i = 1 \tag{4}$$

The first constraint defines a unit hypercube in n-dimensional space; the second defines a hyperplane. Their intersection forms a finite subset known as the "$(n-1)$-simplex", denoted Δ^{n-1}. "Uniformly sampling an n-dimensional distribution" thus means uniformly sampling a point on this simplexmuch more intuitive.

Why use the triangle symbol Δ for simplex? When $n = 3$, the simplex is an equilateral triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$. When $n = 4$, it becomes a regular tetrahedron in 3D space. Indeed, geometrically, a simplex is the high-dimensional generalization of an equilateral trianglehence the use of Δ.

5 Limit Theorem

Now consider computing the entropy probability density, formally written as:

$$\rho(h) = \frac{\int_{\boldsymbol{p} \in \Delta^{n-1}} \delta(H(\boldsymbol{p}) - h) d\boldsymbol{p}}{\int_{\boldsymbol{p} \in \Delta^{n-1}} d\boldsymbol{p}} \tag{5}$$

where $\delta(\cdot)$ is the Dirac delta function. Physically, this "counts" distributions satisfying $H(\boldsymbol{p}) = h$, normalized by the total number. However, this is merely a formal solution; actual analytical computation is nearly impossible.

Given n is typically large in practice (hundreds or thousands), we invoke the Central Limit Theorem to approximate $\rho(h)$ as a normal distribution. A normal distribution is determined by its mean and variance, so we only need to compute the expected entropy and second moment: $\mathbb{E}_{\boldsymbol{p}\sim\Delta^{n-1}}[H(\boldsymbol{p})]$ and $\mathbb{E}_{\boldsymbol{p}\sim\Delta^{n-1}}[H(\boldsymbol{p})^2]$. If we only seek the peak, the mean suffices, as the mode equals the mean for normal distributions.

Computing $\mathbb{E}_{\boldsymbol{p}\sim\Delta^{n-1}}[H(\boldsymbol{p})]$ requires sampling from Δ^{n-1}. While conceptually clear, implementation is non-trivial. Readers familiar with Dirichlet distributions may use known results, but we avoid introducing that here to lower the barrier.

6 Sampling Transformation

Fortunately, uniform sampling from a simplex can be transformed into independent sampling from the exponential distribution Exp(1):

$$\boldsymbol{p} \sim \Delta^{n-1} \quad \Leftrightarrow \quad p_i = \frac{x_i}{\sum_{j=1}^n x_j}, \quad x_1, x_2, \cdots, x_n \sim \text{Exp}(1) \tag{6}$$

Sampling n i.i.d. values from Exp(1) and normalizing yields a uniform sample from the simplex. How to understand this?

A rigorous proof is possible, but here we offer an analogy. Consider uniformly sampling an n-dimensional unit vector: sample n values $x_i \sim \mathcal{N}(0,1)$ and return $\boldsymbol{x}/\|\boldsymbol{x}\|$. This works because the joint density is proportional to $e^{-\|\boldsymbol{x}\|^2/2}$, which is rotationally invarianthence uniform over directions.

A unit vector satisfies "sum of squares equals 1", while our distributions satisfy "sum equals 1". Probability distributions are thus like "unit vectors" under a norm based on the "sum". The exponential distribution Exp(1) has density e^{-x}; n samples give joint density $e^{-\sum x_i}$, depending only on the sum norm. Hence, this sampling is uniform over the new "directions".

7 Field of Averages

We now proceed to calculation. Under this reparameterization, the entropy becomes:

$$H(\boldsymbol{p}) = \log \sum_{i=1}^n x_i - \frac{\sum_{i=1}^n x_i \log x_i}{\sum_{i=1}^n x_i} \tag{7}$$

Taking expectations yields a complex integral. Though solvable analytically with advanced techniques, we instead apply a mean-field approximation:

$$\begin{aligned} \mathbb{E}\left[\log \sum_{i=1}^n x_i - \frac{\sum_{i=1}^n x_i \log x_i}{\sum_{i=1}^n x_i}\right] &\approx \log \sum_{i=1}^n \mathbb{E}[x_i] - \frac{\sum_{i=1}^n \mathbb{E}[x_i \log x_i]}{\sum_{i=1}^n \mathbb{E}[x_i]} \\ &= \log n - (1-\gamma) \end{aligned} \tag{8}$$

where γ is the Euler-Mascheroni constant. Here $\mathbb{E}[x_i] = 1$ is clear; $\mathbb{E}[x_i \log x_i] = 1-\gamma$ is classical, related to derivatives of the Gamma function. Unknown values can be computed numerically (e.g., in Mathematica).

The maximum entropy for n-dimensional distributions is $\log n$. The correction term $-(1-\gamma) \approx -0.42278$ is crucial: it suggests the peak entropy density lies about 0.42 below the maximum. For matrices, this implies greater expressive power near this entropy, not maximal entropy. If singular value entropy must be fixed, this value is a principled choice.

8 General Result

Using similar reasoning, we compute the expected Rényi entropy:

$$\begin{aligned}\mathbb{E}[H_q(\boldsymbol{p})] &= \frac{1}{q-1}\mathbb{E}\left[q\log\sum_{i=1}^{n}x_i - \log\sum_{i=1}^{n}x_i^q\right] \\ &\approx \frac{1}{q-1}\left[q\log\sum_{i=1}^{n}\mathbb{E}[x_i] - \log\sum_{i=1}^{n}\mathbb{E}[x_i^q]\right] \\ &= \frac{1}{q-1}\left[q\log n - \log(n\Gamma(q+1))\right] \\ &= \log n - \frac{\log\Gamma(q+1)}{q-1}\end{aligned} \tag{9}$$

As $q \to 1$, the limit is $\log n - (1-\gamma)$, consistent with the previous section. If q is an integer greater than 1, we have:

$$\mathbb{E}[H_q(\boldsymbol{p})] \approx \log n - \frac{\log q!}{q-1} \tag{10}$$

For $q = 2$, the result is $\log(n/2)$the maximum entropy of an $n/2$-dimensional distribution. This suggests optimal singular value entropy only fills about half the dimensions; full saturation reduces diversity.

9 Complete Analysis

The above focused on estimating the mean, sufficient for our goal. For completeness, this section estimates variance and discusses asymptotic normality.

Rewriting $H_q(\boldsymbol{p})$:

$$H_q(\boldsymbol{p}) = \log n + \frac{1}{q-1}\left[q\log\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i}_{u} - \log\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i^q}_{v}\right] \tag{11}$$

For $q \neq 1$, u and v are distinct statistics. The Central Limit Theorem implies (u, v) asymptotically follows a bivariate normal distribution as $n \to \infty$. Under $x_i \sim \text{Exp}(1)$, u has mean 1 and variance $1/n$, while v has mean $\Gamma(q+1)$ and variance $[\Gamma(2q+1) - \Gamma(q+1)^2]/n$.

For moderate q and large n, u and v concentrate near their means. A first-order Taylor expansion gives:

$$H_q(\boldsymbol{p}) \approx \log n + \frac{1}{q-1}\left[q(u-1) - \left(\log\mu_v + \frac{v-\Gamma(q+1)}{\Gamma(q+1)}\right)\right] \tag{12}$$

Taking expectation recovers Eq. (9). For more accuracy, higher-order terms can be included.

Since $H_q(\boldsymbol{p})$ is approximately linear in u and v, its variance is:

$$\mathbb{V}\text{ar}[H_q(\boldsymbol{p})] \approx \mathbb{V}\text{ar}\left[\frac{qu}{q-1}\right] + \mathbb{V}\text{ar}\left[\frac{v}{(q-1)\Gamma(q+1)}\right] - 2\mathbb{C}\text{ov}\left[\frac{qu}{q-1}, \frac{v}{(q-1)\Gamma(q+1)}\right] \tag{13}$$

Note u and v are generally not independent; the covariance term must be retained:

$$\mathbb{C}\text{ov}[u, v] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}\text{ov}[x_i, x_j^q] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{C}\text{ov}[x_i, x_i^q] \tag{14}$$

The second equality holds because x_i, x_j are independent for $i \neq j$. We compute $\mathbb{C}\mathrm{ov}[x_i, x_i^q] = \mathbb{E}[x_i^{q+1}] - \mathbb{E}[x_i]\mathbb{E}[x_i^q] = \Gamma(q+2) - \Gamma(q+1) = q\Gamma(q+1)$. Substituting:

$$\mathbb{V}\mathrm{ar}[H_q(\boldsymbol{p})] \approx \frac{1}{n}\left[\frac{\Gamma(2q+1)}{(q-1)^2\Gamma(q+1)^2} - \frac{q^2+1}{(q-1)^2}\right] \tag{15}$$

For $q = 2$, this gives $1/n$; for $q \to 1$, it gives $(\pi^2/3 - 3)/n$.

In summary, while the Central Limit Theorem does not directly imply asymptotic normality of h, it guarantees it for intermediate variables u, v, from which we estimate h's mean and variance. The variance is $\mathcal{O}(1/n)$, shrinking with n, justifying locating the peak via the mean.

10 Summary

This article reframed the question "Is higher singular value entropy better?" into a quantifiable mathematical problem. The core hypothesis: fixed singular value entropy corresponds to remaining degrees of freedom, which reflect expressive capacity. We then identified the point of maximum entropy density as optimal, transforming and approximating to derive a reference value.

While the result itself may not be paramount, the article's significance lies in providing a framework for analyzing such questions.