The Geometry of the Three Reductions
companion to *Four Pictures of One Matmul*

Japjot Singh

typeset from the session discussion of October 28, 2026

Setting, as in the main note: $W \in \mathbb{R}^{T\times T}$ is the softmax attention matrix (row i is reader i's recipe), $V \in \mathbb{R}^{T\times d}$ holds the payloads (row k is source k's), and $B[i,k,j] = W[i,k]\,V[k,j]$ is the term block. The block admits three reductions; each has a geometric reading, and the readings are not equally meaningful.

1 Two spaces

W's rows and V's rows are both vectors, in different spaces with different geometry:

- a **recipe** $W[i,:]$ lives *on the probability simplex* in $\mathbb{R}^T$: entries nonnegative, summing to 1. Its coordinates are indexed by tokens.
- a **payload** $V[k,:]$ lives in unconstrained *feature space* $\mathbb{R}^d$. Its coordinates are feature dimensions, and no direction in this space is privileged.

This settles "W is scores, V is vectors": both are vectors; the simplex constraint is what makes W's rows feel like scores. The roles (coefficient vs. thing-being-combined) are assigned by each picture of the matmul, not intrinsic to the matrices: the row picture spends W's entries as coefficients on V's rows; the column picture does the reverse.

2 The three reductions

Sum over k (the contraction): a center of mass. This is the one reduction the model intends: $\mathrm{ctx}[i,:] = \mathbb{E}_{k\sim p_i} V[k,:]$ with $p_i = W[i,:]$, the recipe-weighted center of mass of the payload points, confined to their convex hull (derived and drawn in the main note; not repeated here).

Sum over i (the readers): a simplex centroid. Stack all T recipes as T points on the simplex (Figure 1). The column-mass vector

$$m[k] \;=\; \sum_i W[i,k]$$

is exactly T times the *centroid* of that point cloud: coordinate k of the average reader's distribution, scaled by T. Probabilistically, if every reader sampled one source from its row distribution,

$m[k]$ is source k's **expected fetch count**. The reduction

$$B.\text{sum}(\text{dim=0})[k,:] \;=\; m[k]\; V[k,:]$$

is therefore *payloads weighted by audience*: each source's vector, scaled by how much attention it received in total. Rows of W sum to 1, so the whole budget is fixed at T and column masses genuinely compete for it; comparing them is legitimate arithmetic. Two caveats before reading them as importance:

1. **Positional availability.** Under the causal mask, source k is visible only to readers $i \geq k$, so early tokens have more possible readers: raw column mass confounds popularity with position.
2. **Attention sinks.** Token 0 tends to hoard column mass: a reader that wants to attend to nothing still must spend a budget of 1, and the leftover lands on the first token(s) (the StreamingLLM observation; module M6 material). More generally, whether attention mass measures "importance" at all is contested in the interpretability literature: treat m as a diagnostic, not ground truth.

Sum over j (the features): a projection onto an arbitrary axis. For each (i,k) pair,

$$B.\text{sum}(\text{dim=2})[i,k] \;=\; W[i,k]\sum_j V[k,j] \;=\; W[i,k]\,\big\langle V[k,:],\,1\big\rangle,$$

and $\langle V[k,:],1\rangle$ is (up to $\sqrt{d}$) the projection of payload k onto the all-ones direction (Figure 2). In feature space the all-ones direction is *not special*: rotate the basis and it points somewhere else. So this reduction scales attention weights by a basis-dependent coordinate sum: legal arithmetic, no meaning. The inability to find words for it is correct perception, not confusion.

The general rule: summing a free index collapses the factor that owns it into the shape of the other. Summing i (owned by W) leaves V's vector structure intact, scaled; summing j (owned by V) leaves only W, scaled by coordinate junk. Only summing the shared k performs the gather.

The SQL hook (mine)

Contraction in one line of SQL: `SELECT i, j, SUM(W*V) GROUP BY i, j`. Free indices are the GROUP BY keys; the contracted index is the summed column. The k-sum is the only reduction the model intends; the other two are queries nobody asked for.

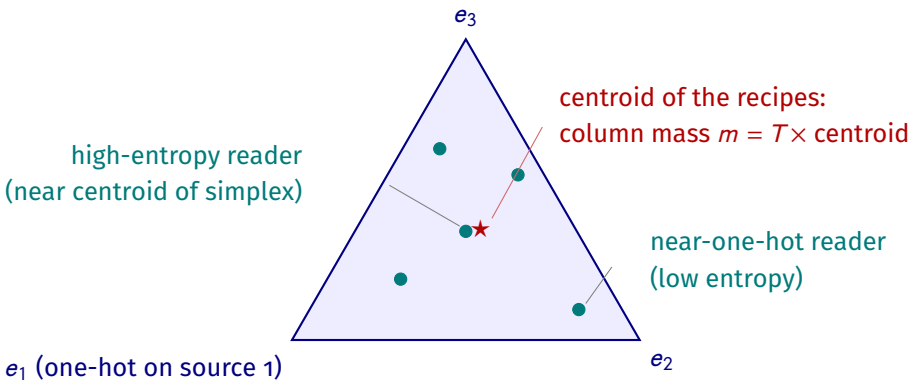


Figure 1: Sum over readers, on the simplex. Each teal point is one reader's recipe $W[i,:]$ (a probability vector over T=3 sources). The vertices are one-hot recipes; points near the middle are high-entropy readers. The column-mass vector is T times the centroid ($\star$) of the cloud: coordinate k is source k's expected fetch count.

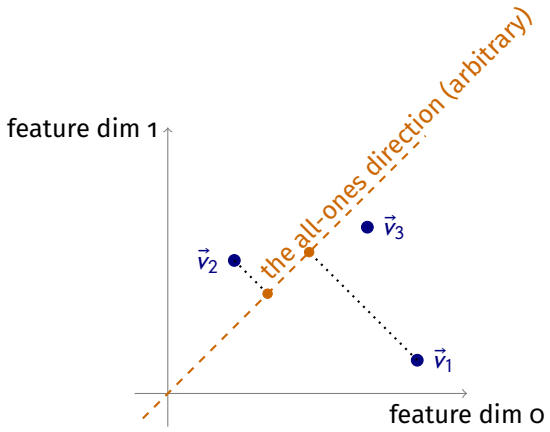


Figure 2: Sum over features: each payload collapses to its component along the all-ones line (orange feet). Nothing in feature space privileges that line; under a basis rotation it moves. The reduction is legal and meaningless.