

Where Did the Context Matrix Go?

Gauge freedom, induced metrics, and what W is actually carrying

James Pustejovsky

Brandeis University
Department of Computer Science

CS 135: Computational Semantics

Fall 2026

The Question

SGNS trains two tables, W and C . At the end we keep W , call it the word embeddings, publish it, and never mention C again.

The usual justification is that C was scaffolding. That is not an argument, and the thing it is standing in for turns out to be worth an hour.

Three questions

- ① In what sense, if any, is C 's information *in* W ?
- ② What exactly are we assuming when we compute $\cos(\mathbf{w}_i, \mathbf{w}_j)$ and call it similarity?
- ③ Is there a version of the model in which discarding C is provably free?

The answer to the third is yes, and constructing it is the substance of today. The answer to the second is a bilinear form nobody writes down.

The Score Is a Bilinear Form

Everything SGNS computes passes through one expression:

$$\text{score}(w, c) = \mathbf{w}_w \cdot \mathbf{c}_c = \mathbf{w}_w^\top \mathbf{c}_c$$

Written with the middle made visible:

$$\text{score}(w, c) = \mathbf{w}_w^\top I \mathbf{c}_c$$

That I is a choice, and it was never made explicitly. A bilinear form $B(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top G \mathbf{y}$ needs a matrix G in the middle, and taking $G = I$ says: the coordinates are orthonormal, and each contributes equally.

In Session 6B you met bilinear forms as relations: $\mathbf{x}^\top M \mathbf{y}$ asked whether x stands in relation M to y , and M carried the content. Here the content is supposed to be in the vectors and the middle is supposed to be inert. Whether that is true is the question.

What Training Actually Pinned Down

The objective only ever sees products $\mathbf{w}_w \cdot \mathbf{c}_c$. Collecting them:

$$WC^\top \approx M, \quad M_{wc} = \text{PMI}(w, c) - \log k$$

Determined by training	the matrix product $WC^\top$
Not determined	W and C individually

That second row is not a statement about optimization being imperfect. It is exact, and it has a name.

The factorization has a continuous symmetry, and every quantity we compute from W alone had better be checked against it.

A Symmetry in the Factorization

Take any invertible $A \in \mathbb{R}^{d \times d}$ and set

$$W' = WA, \quad C' = CA^{-\top}$$

Then

$$W'C'^{\top} = (WA)(CA^{-\top})^{\top} = WAA^{-1}C^{\top} = WC^{\top}$$

The pair (W, C) and the pair $(WA, CA^{-\top})$ are **indistinguishable to the objective**. Every value of the loss, every gradient, every prediction is identical. There are d^2 continuous directions of freedom – the group $\text{GL}(d, \mathbb{R})$ acts on the solution.

So “the” embedding produced by SGNS is not a matrix. It is an equivalence class of matrices, and which representative you receive is decided by initialization and by the path the optimizer happened to take.

Verified, on the Trained Model

Take the model from last session and apply a random invertible A :

$\max_{ij} (WC^\top)_{ij} - (W'C'^\top)_{ij} $	2.7×10^{-15}
$\text{mean } \cos(\mathbf{w}'_i, \mathbf{w}'_j) - \cos(\mathbf{w}_i, \mathbf{w}_j) $	0.0972

The reconstruction is preserved to machine precision.

The word-to-word cosines move by about 0.1 on average, and nothing in the training objective prefers one answer to the other.

Cosine similarity between word vectors is therefore not a property of the model. It is a property of the *representative* of the model that you happen to be holding.

Which Quantity Is the Real One?

If $\cos(\mathbf{w}_i, \mathbf{w}_j)$ is gauge-dependent, we should ask what the *gauge-independent* statement about two words would be.

The model's claim about word i is its whole row of the reconstruction:

$$\mathbf{m}_i = \text{row } i \text{ of } WC^\top = (\mathbf{w}_i \cdot \mathbf{c}_1, \dots, \mathbf{w}_i \cdot \mathbf{c}_{|V|})$$

the predicted association of w_i with every context. *That* is invariant.

The true similarity

$$\langle \mathbf{m}_i, \mathbf{m}_j \rangle = (WC^\top CW^\top)_{ij} = \mathbf{w}_i^\top (\textcolor{red}{C}^\top \textcolor{red}{C}) \mathbf{w}_j$$

There is the bilinear form. Comparing two words is $\mathbf{w}_i^\top G \mathbf{w}_j$ with $G = C^\top C$, a $d \times d$ matrix built entirely out of the context table.

So What Does Discarding C Assume?

true: $\mathbf{w}_i^\top (C^\top C) \mathbf{w}_j$

what we compute: $\mathbf{w}_i^\top I \mathbf{w}_j$

The assumption, stated

Using W with a plain dot product is exactly correct when

$$C^\top C = \lambda I$$

for some scalar λ – that is, when the columns of C are orthogonal and of equal length. Then the two forms agree up to an overall scale, which cosine removes.

This is a checkable condition on a $d \times d$ matrix, not a philosophical position. So check it.

It Is Not True

$C^\top C$ for the trained model, $d = 12$:

diagonal entries, mean	4.845
diagonal entries, standard deviation	3.921
off-diagonal, mean absolute value	1.656
off-diagonal, largest absolute value	10.349
$\ C^\top C - \lambda I\ _F / \ C^\top C\ _F$	0.891

The columns are neither equal in length nor close to orthogonal. The largest off-diagonal entry exceeds twice the mean diagonal.

Nearly ninety percent of $C^\top C$, measured in Frobenius norm, is the part that discarding C throws away.

And Yet It Mostly Works

Compare each candidate embedding against the true row geometry of the reconstruction:

embedding used	correlation	mean error
W alone, what word2vec ships	0.9826	0.2681
$W + C$	0.9840	0.8682
the reconstruction itself	1.0000	0.0000

Correlation 0.98 says the *ordering* of similarities largely survives, which is why nobody notices. Mean error 0.27 on a quantity bounded by 1 says the *values* do not.

Every reported cosine from a word2vec model carries an error of roughly that size, in a direction determined by a matrix that was deleted.

The Interesting Part: You Can Make It True

$C^\top C$ is symmetric and positive definite, so it has a Cholesky factorization

$$C^\top C = LL^\top$$

Use L as the gauge transformation:

$$\boxed{W' = WL, \quad C' = CL^{-\top}}$$

Two things follow immediately. The reconstruction is untouched,

$$W'C'^\top = WLL^{-1}C^\top = WC^\top$$

and the new context table is orthonormal,

$$C'^\top C' = L^{-1}(C^\top C)L^{-\top} = L^{-1}LL^\top L^{-\top} = I$$

$\max_{ij} W' C'^{\top} - W C^{\top} $	1.8×10^{-15}
$\max_{ij} C'^{\top} C' - I $	6.7×10^{-9}

What just happened

The model is unchanged: same predictions, same loss, same every observable. But the metric that was sitting inside C has been moved into W . In this basis C is a plain orthonormal frame carrying no information about scale or correlation at all.

Now the sentence is true. C can be discarded because its content has been absorbed into W' – not as a figure of speech, but as a matrix identity.

The Absorbed Embedding

embedding used	correlation	mean error	mean cosine
W alone	0.9826	0.2681	0.536
$W + C$	0.9840	0.8682	-0.064
$W' = WL$	1.0000	0.0000	0.804
the reconstruction itself	1.0000	0.0000	0.804

W' reproduces the true geometry exactly, in d dimensions, with an ordinary dot product. No approximation is involved.

It is a one-line correction – a Cholesky factorization and a matrix multiply – and essentially nobody applies it, because essentially nobody publishes C .

Note the last column. The corrected space is *more* anisotropic than the raw one, not less.

Does It Change the Answer You Would Report?

Nearest neighbors, before and after absorbing C :

	from W	from $W' = WL$
coffee	tea 0.98, brew 0.96, morning 0.92	tea 1.00, morning 1.00, brew 1.00
laptop	screen 0.96, boot 0.93, charger 0.92	boot 1.00, screen 1.00, charger 1.00
river	boat 0.99, water 0.98, fish 0.96	boat 1.00, water 1.00, fish 1.00
battery	laptop 0.91, screen 0.89, boot 0.85	boot 1.00, laptop 1.00, screen 1.00

The membership of the top three is stable. The *ranking* shifts in two of four cases, and the values change character completely.

Read the left column again. Those graded values – 0.98, 0.96, 0.92 – look like the model making fine distinctions among the beverages. In the corrected metric they are all 1.00.

The gradation was not in the data. It was the discarded metric, showing up as apparent structure.

This corpus was built from three clean topics, so exact interchangeability is the correct answer here. On real text the effect is smaller and in the same direction.

Why the SVD Never Had This Problem

Run the same check on the truncated SVD, $M_k = U_k \Sigma_k V_k^\top$, taking $W = U_k \Sigma_k$ and $C = V_k$:

k	$V_k^\top V_k = I?$	cos from $U_k \Sigma_k$ vs. from M_k
2	exactly	correlation 1.000000
6	exactly	correlation 1.000000
12	exactly	correlation 1.000000

The SVD delivers its factorization already in the whitened gauge. V is orthonormal by construction, so $C^\top C = I$ on the nose and discarding V is free – exactly, at every k .

This is what the orthogonality of U and V was buying all along, and it is why Session 8's only open question was where to put Σ . The gauge had already been fixed; all that remained was the scale.

The α Exponent, Seen Properly

Session 8 asked where the scale should go: $W = U_k \Sigma_k^\alpha$, $C = V_k \Sigma_k^{1-\alpha}$.

α	$C^\top C$	status
1	I	whitened; discarding C is exact
1/2	Σ_k	diagonal but unequal
0	Σ_k^2	diagonal, badly unequal

So α is a one-parameter slice through the $\text{GL}(d)$ gauge group – the diagonal part of it. The SVD's orthogonality removes every off-diagonal degree of freedom in advance and leaves exactly d knobs, which α then ties together into one.

SGNS has all d^2 knobs loose, and sets them by accident.

Anisotropy Is a Symptom of This

Contextual and static embeddings are both reported as strongly anisotropic: random word pairs have high average cosine.

mean cosine, W as shipped	0.536
mean cosine, W' in the whitened gauge	0.804
mean cosine, the reconstruction itself	0.804

Two readings are available, and they are not the same. Either the model genuinely thinks these words are all rather similar, in which case the anisotropy is a fact about the corpus; or the number depends on which representative of the equivalence class you measured, in which case it is partly an artifact.

Here it is both. The raw W understates it and the corrected W' matches the reconstruction, so the geometry was being misreported – but not in the direction the “anisotropy is a defect” literature assumes. Which is a reason to be careful with post-hoc fixes that change the gauge without saying so.

What This Was Really About

A bilinear form	comparing two things needs $\mathbf{x}^\top G \mathbf{y}$; the middle matrix is the metric, and it is never optional
Learned from data	SGNS learns $G = C^\top C$ without ever naming it, then deletes it
Gauge freedom	the data fixes $WC^\top$, not W ; the metric and the coordinates trade off against each other
Absorption	a change of basis can move the metric into the coordinates, and then the coordinates are enough

A vector is not a meaning. A vector *plus a form* is a meaning, and the form has to come from somewhere. Distributional semantics gets one by accident and throws it away.

Two Other Ways to Get the Metric

Distributional	learn G implicitly as $C^\top C$, discard it, and hope $G \propto I$
VSA	<i>assume</i> it. Random high-dimensional vectors are quasi-orthogonal by construction, so $G \approx I$ holds by design rather than by luck, and the margin improves as $1/\sqrt{d}$
Geometric algebra	<i>declare</i> it. The signature $\text{Cl}(p, q, r)$ fixes the quadratic form before any computation, and every product is defined relative to it

Three positions on one question. The first leaves the metric implicit and unexamined; the second builds a space in which the assumption is safe; the third writes it down and makes it part of the object language.

When we get to $\text{Cl}(3, 0)$ and ask why the signature must be chosen before anything else is defined, this is the answer: the signature *is* G , and today showed what becomes of a representation that has none.

Summary

Point	Detail
The score	$\mathbf{w}^\top I \mathbf{c}$: a bilinear form whose middle matrix was chosen silently
What training fixes	$WC^\top$, and nothing finer
Gauge freedom	$(W, C) \rightarrow (WA, CA^{-\top})$ for any invertible A : d^2 free directions
Consequence	word cosines are not a property of the model; they moved 0.097 under a random A
The true form	$\mathbf{w}_i^\top (C^\top C) \mathbf{w}_j$ – the metric is built from the context table
Discarding C assumes	$C^\top C \propto I$, which is false here by 89% in Frobenius norm
Cost of the assumption	correlation 0.98 with the truth, mean cosine error 0.27
The fix	$C^\top C = LL^\top$; set $W' = WL$, $C' = CL^{-\top}$; then $C'^\top C' = I$ exactly
After the fix	W' alone reproduces the true geometry with correlation 1.0000
Why the SVD is clean	V is orthonormal by construction, so it ships in the whitened gauge
α , revisited	the diagonal slice of the gauge group; SGNS leaves all d^2 loose
Anisotropy	partly a gauge artifact, and not in the direction usually assumed
The lesson	a vector plus a form is a meaning; the form must come from somewhere
VSA	assumes $G \approx I$ by construction
Geometric algebra	declares G as a signature, before anything else is defined

Exercises

- 1 Show that if A is orthogonal, the gauge transformation $(W, C) \rightarrow (WA, CA^{-\top})$ leaves *all* word cosines unchanged. Which subgroup of $GL(d)$ is therefore harmless, and how large is it compared with the whole group?
- 2 The Cholesky factor L is not the only matrix with $LL^{\top} = C^{\top}C$. Give another, and say what the corresponding W' differs by. Does the choice affect any cosine?
- 3 We used $C^{\top}C$ to compare words. Write down the corresponding form for comparing two *contexts*, and say which table supplies the metric in that case.
- 4 $W + C$ scored a slightly higher correlation than W alone but a much larger mean error, and its mean cosine was -0.064 against the true 0.804 . Explain what addition is doing to the geometry, and why correlation and error disagree.
- 5 **Harder.** SGNS with tied tables, $C = W$, is sometimes proposed. What does tying do to the gauge freedom, what does it force G to be, and what does it cost?

Questions?

A vector plus a form is a meaning.
Next: what happens when you declare the form.