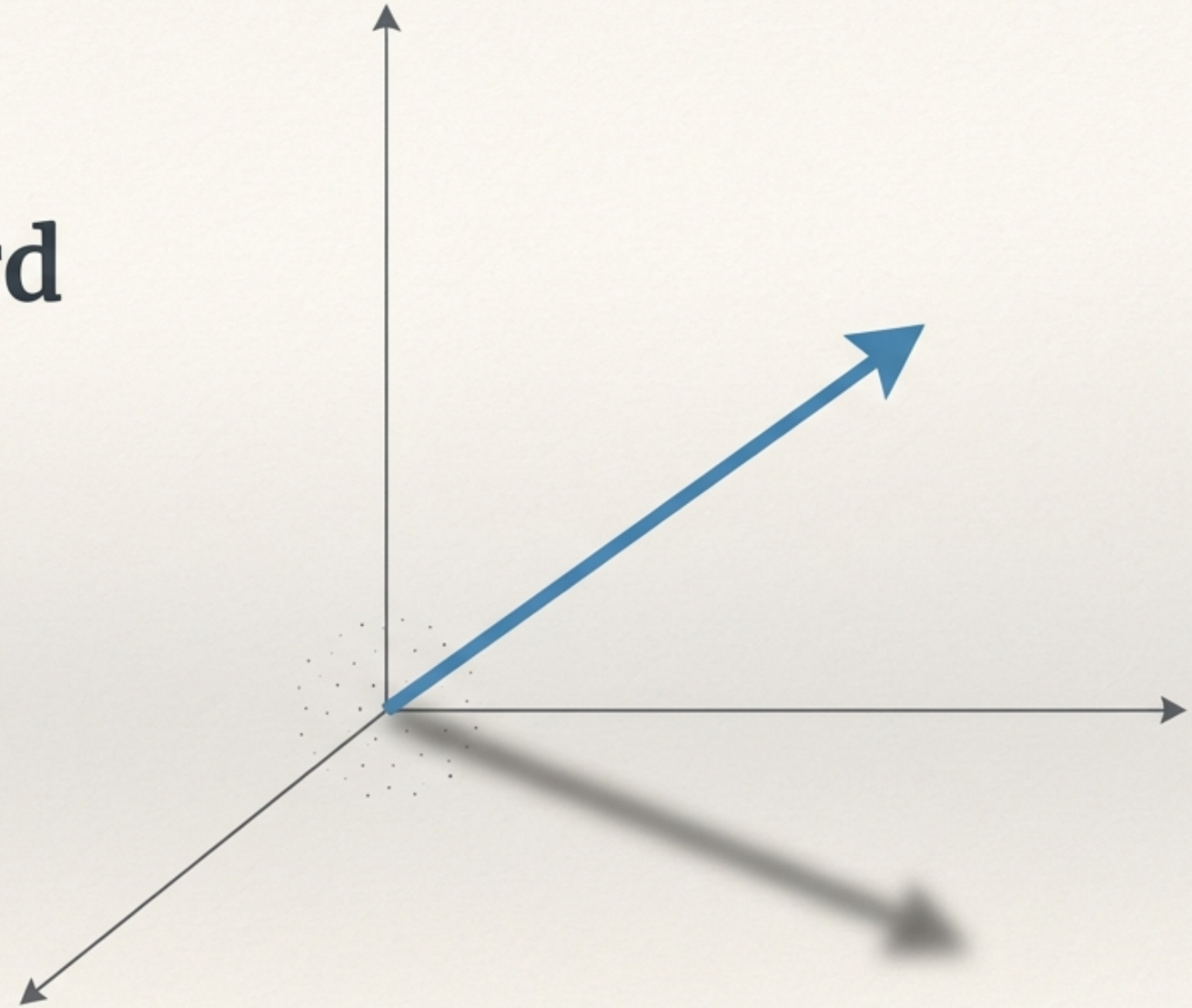
A Rigorous Treatment of Word Embeddings

From Distributional Theory to Geometric Semantics



The Catastrophic Sparsity of One-Hot Vectors

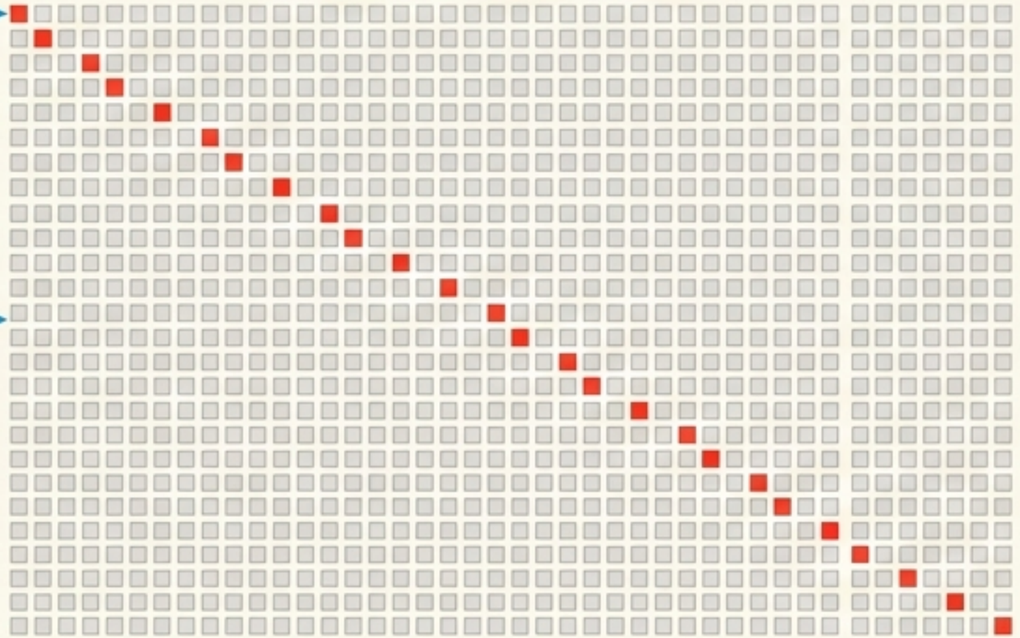
$$\phi: V \rightarrow \mathbb{R}^d$$

V: Discrete vocabulary of size $V \in [10^4, 10^6]$ $\mathbb{R}^d$: Dense target space $d \ll V$

$$\mathbf{o}_i^T \mathbf{o}_j = \delta_{ij}, \quad \delta_{ij} = 0 \text{ for all } i \neq j$$

Core Insight: One-hot encoding enforces **false orthogonality**. The inner product carries **zero information** about linguistic relatedness, obliterating any semantic or syntactic similarity signal.

One-Hot Representation



Dense Continuous Vector

The Distributional Hypothesis

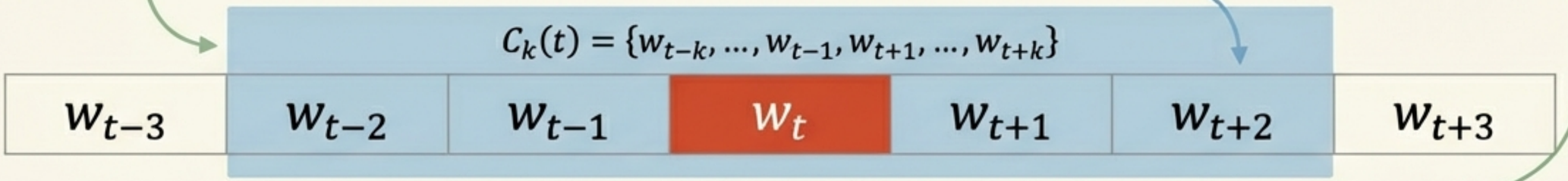
Axiom Box

A word is characterized by the company it keeps. — J.R. Firth (1957)

Meaning is defined by conditional distribution:

$$D_{KL}\left(P(\text{context} \mid w_i) \parallel P(\text{context} \mid w_j)\right) \approx 0$$

When **context distributions overlap**, KL divergence approaches zero, meaning words occupy similar semantic roles.



The Bridge to Geometry:

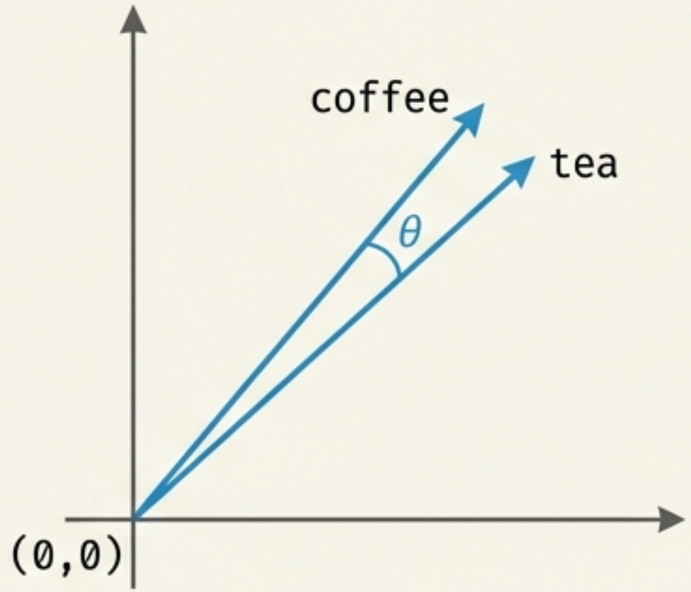
$$\cos\left(\phi(w_i), \phi(w_j)\right) \approx f\left(\texttt{distributional\_overlap}(w_i, w_j)\right)$$

Geometric encoding must mirror context distribution.

Vector Semantics and the Geometry of Meaning

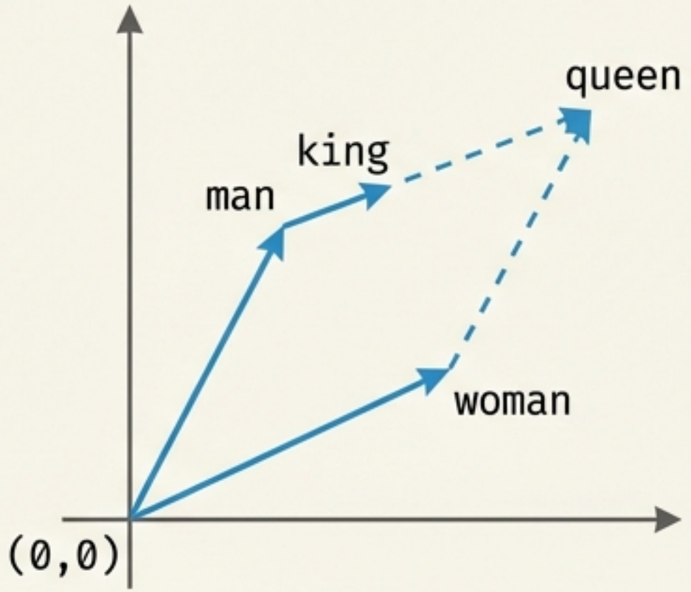
Similarity (Angular Proximity)

$$\cos\theta_{ij} = \frac{\mathbf{e}_i^T \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}$$



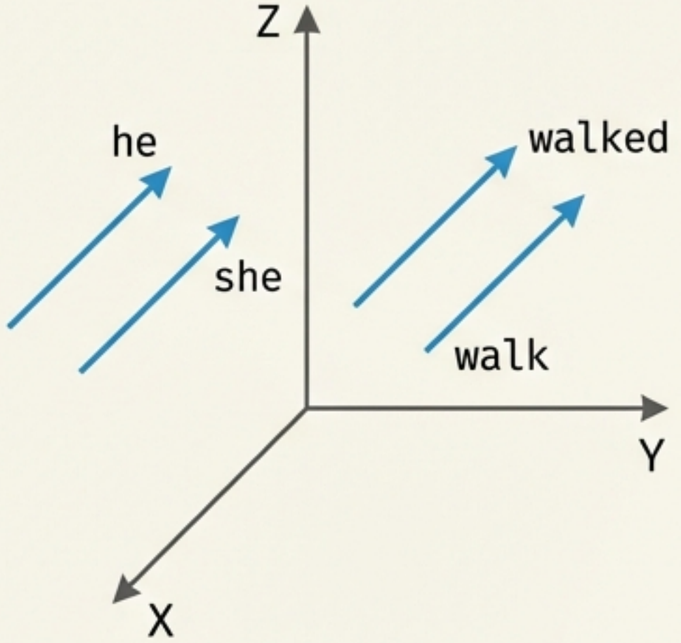
Analogy (Parallelogram Structure)

$$\mathbf{e}\_king - \mathbf{e}\_man + \mathbf{e}\_woman \approx \mathbf{e}\_queen$$



Attributes (Semantic Axes)

$$\mathbf{v}\_gender \approx \mathbf{e}\_she - \mathbf{e}\_he$$
$$\mathbf{v}\_tense \approx \mathbf{e}\_walked - \mathbf{e}\_walk$$



The Top-Level Taxonomy of Word Embeddings

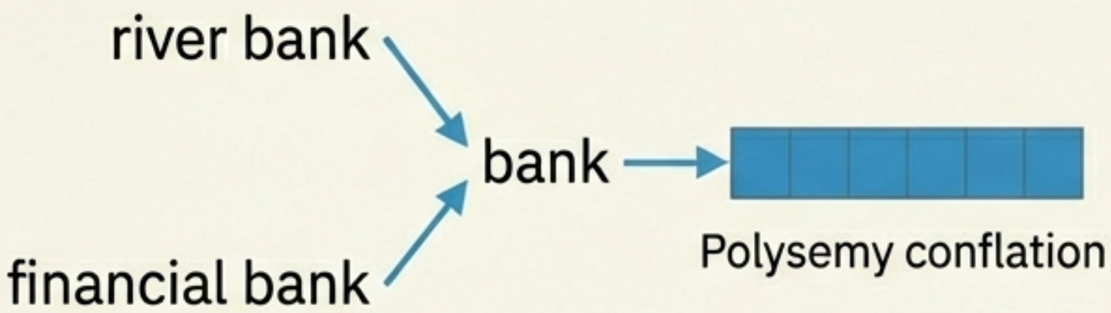
Word Embeddings

Static Embeddings

Assigns a single vector per word type.

$$\phi_{\text{static}}: w \mapsto \mathbf{e}_w \in \mathbb{R}^d$$

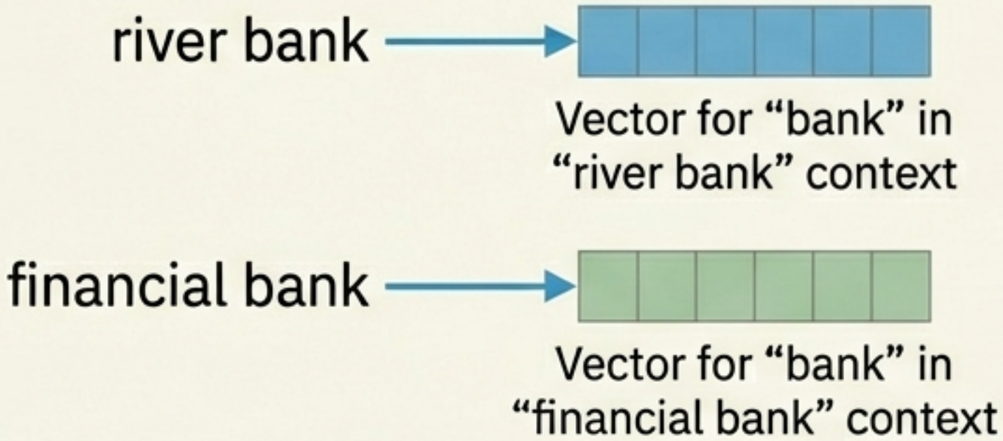
Context-independent.



Contextual Embeddings

Assigns vectors per word token, conditioned on surrounding context.

$$\phi_{\text{contextual}}: (w, C) \mapsto \mathbf{h}_w^{(C)} \in \mathbb{R}^d$$

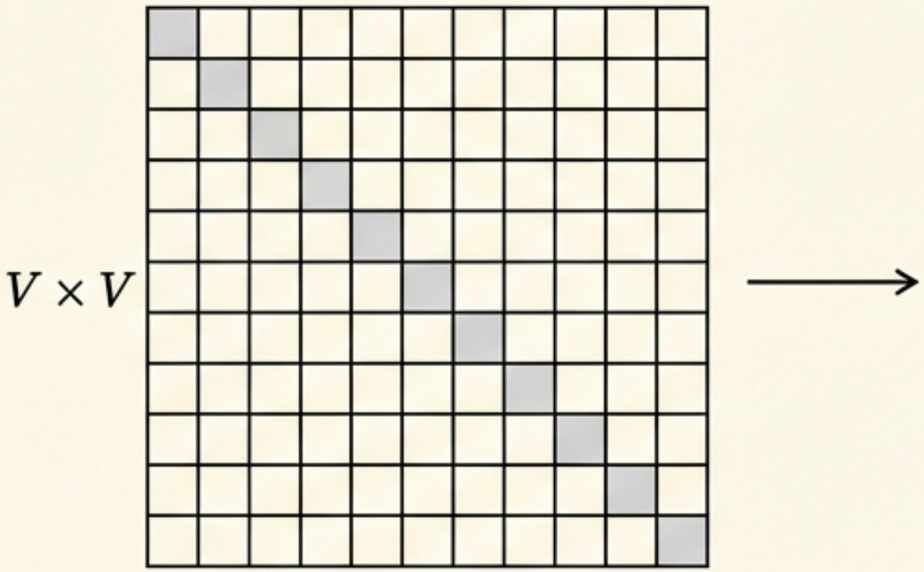


Paradigm 1: Count-Based Methods

Step 1: PPMI Matrix

Raw counts fail due to frequent words.

$$\text{PPMI}(w_i, w_j) = \max(0, \frac{\log_2 P(w_i, w_j)}{P(w_i)P(w_j)})$$

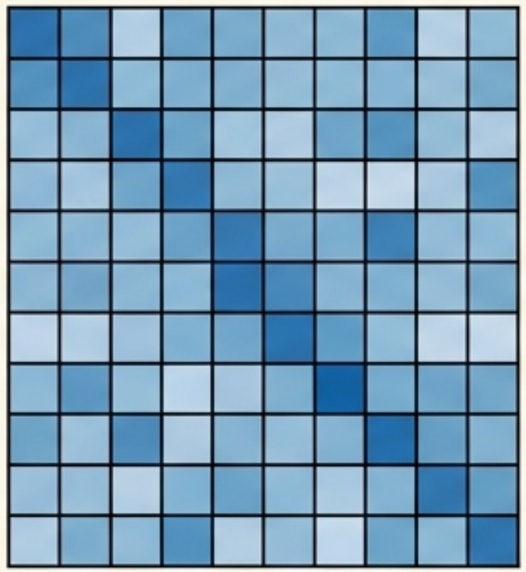


Raw Counts

Step 2: Truncated SVD (LSA)

Compresses the $V \times V$ matrix into d dimensions:

$$M \approx U_d \Sigma_d V_d^T$$

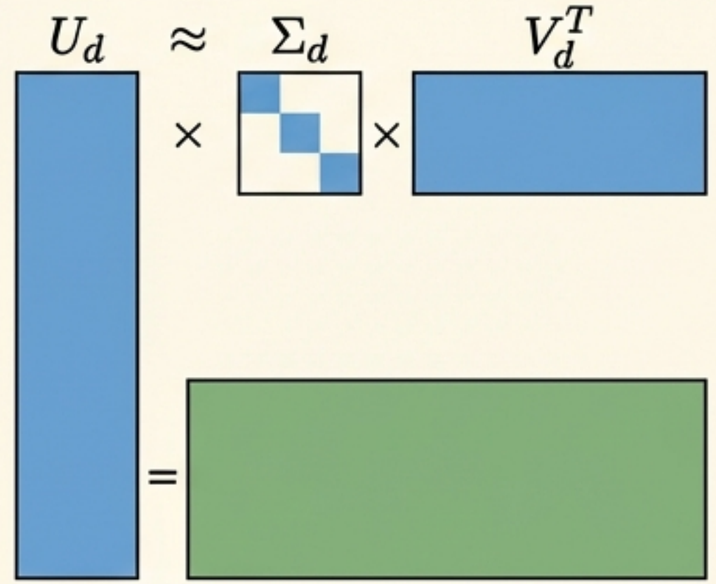


PPMI

Step 3: Optimal Dense Vectors

$$E = U_d \Sigma_d^\alpha$$

$\alpha = 0.5$ balances pure rotation and full magnitude scaling.

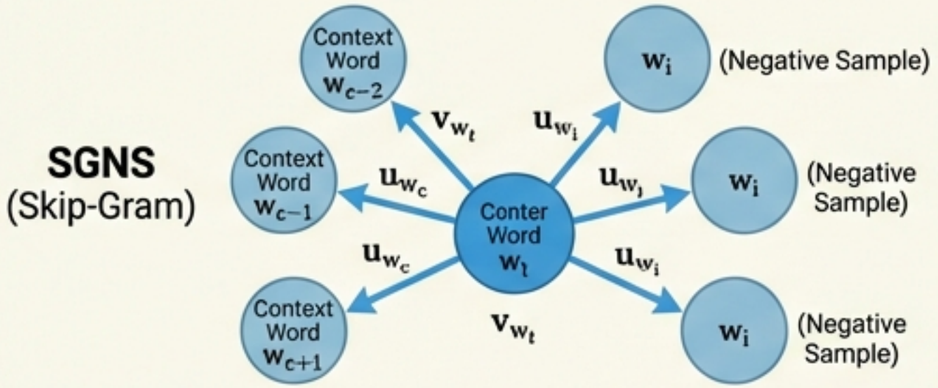


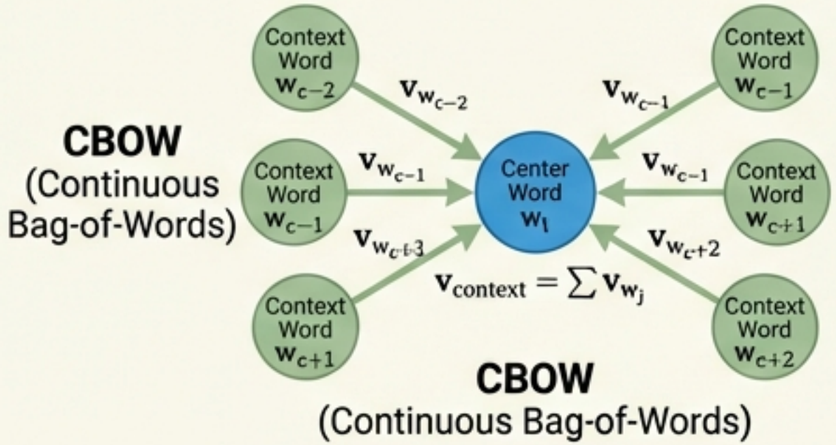
Optimal Dense Vectors E

Paradigm 2: Prediction-Based Methods

The Softmax Bottleneck

Predicting context w_c from center w_t requires a partition function scaling with O(V). Intractable.





Skip-Gram with Negative Sampling (SGNS)

$$\mathcal{L}_{\text{SGNS}} = \log \sigma(\mathbf{u}_{w_c}^{\mathrm{T}} \mathbf{v}_{w_t}) + \sum_{i=1}^{K} E_{w_i} \log \sigma(-\mathbf{u}_{w_i}^{\mathrm{T}} \mathbf{v}_{w_t})$$

$$\mathcal{L}_{\text{SGNS}} = \underbrace{\log \sigma(\mathbf{u}_{w_c}^{\mathrm{T}} \mathbf{v}_{w_t})}_{\text{True Context}} + \underbrace{\sum_{i=1}^{K} E_{w_i} \log \sigma(-\mathbf{u}_{w_i}^{\mathrm{T}} \mathbf{v}_{w_t})}_{\text{Negative Samples}}$$

Binary classification over K negative samples replaces the massive softmax denominator, transforming the intractable O(V) problem into K+I efficient logistic regressions.

Paradigm 3: GloVe and FastText

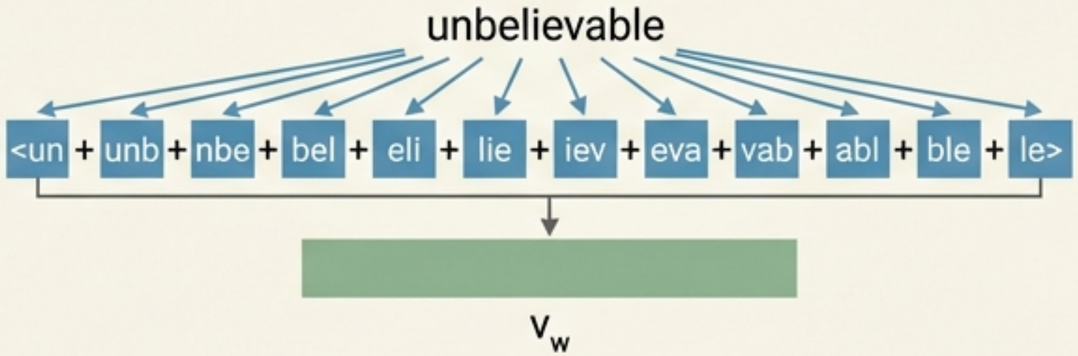
GloVe (Global Vectors)

	P(k|ice)	P(k|steam)
k=solid	↑ Large ↑	↓ Small ↓
k=gas	↓ Small ↓	≈ 1
k=water	≈ 1	≡ 1

Meaning is encoded in co-occurrence ratios. P(k|ice) / P(k|steam) is large for solid, small for gas, ≈ 1 for water.

$$J = \sum f(X_{ij}) \left(w_i^T \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

FastText (Subwords)



Shatters the word boundary to solve the Out-of-Vocabulary (OOV) problem.

Captures morphological regularity (‘un-’, ‘-tion’) and handles morphologically rich languages.

$$\mathbf{v}_w = \frac{1}{|G_w|} \sum_{g \in G_w} \mathbf{z}_g$$

The Grand Unification of Static Embeddings

The Core Revelation (Levy & Goldberg, 2014)

SGNS implicitly factorizes the shifted PMI matrix. All major methods are implicitly or explicitly factorizing variants of the co-occurrence matrix.

$$\mathbf{v}_w^T \mathbf{u}_c \approx \mathrm{PMI}(w, c) - \log K$$

SVD-PPMI	SGNS	GloVe
Factorizes: $\mathrm{PPMI}(w, c)$	Factorizes: $\mathrm{PMI}(w, c) - \log K$	Factorizes: $\log X_{ij}$
Weighting: Uniform	Weighting: Frequency-smoothed	Weighting: $f(X_{ij})$ capped
Optimization: Closed-form SVD	Optimization: SGD	Optimization: AdaGrad

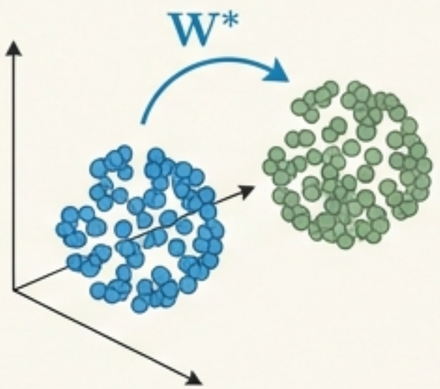
"Hyperparameter tuning matters more than the choice of algorithm."

Applying Embeddings and Transfer Learning

Cross-Lingual Alignment

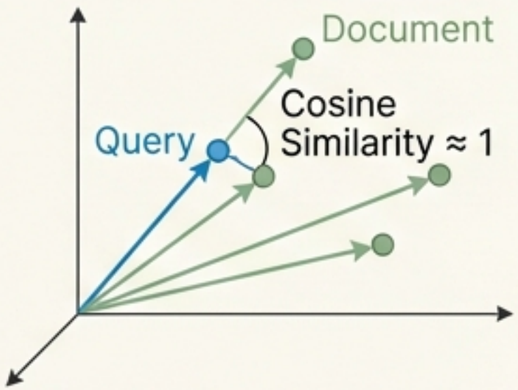
Aligning spaces via Procrustes matrix. Orthogonal rotation preserves monolingual quality.

$$\mathbf{W}^* = \mathbf{V}\mathbf{U}^T$$



Information Retrieval

Query-document matching via cosine similarity.



Transfer Learning Rule of Thumb

When to freeze vs. fine-tune?

Fine-tune only if: $|\mathcal{D}_{\text{task}}| \gg d \cdot \mathbf{V}_{\text{active}}$

Ensure sufficient data exists to avoid destroying the pre-trained geometry.

From Words to Documents

Strategy 1: Simple Averaging

$$\mathbf{d}_{\text{avg}} = \frac{1}{|D|} \sum \mathbf{e}_{\text{w}}$$

Fast, but gives equal weight to uninformative words.

Strategy 2: TF-IDF Weighting

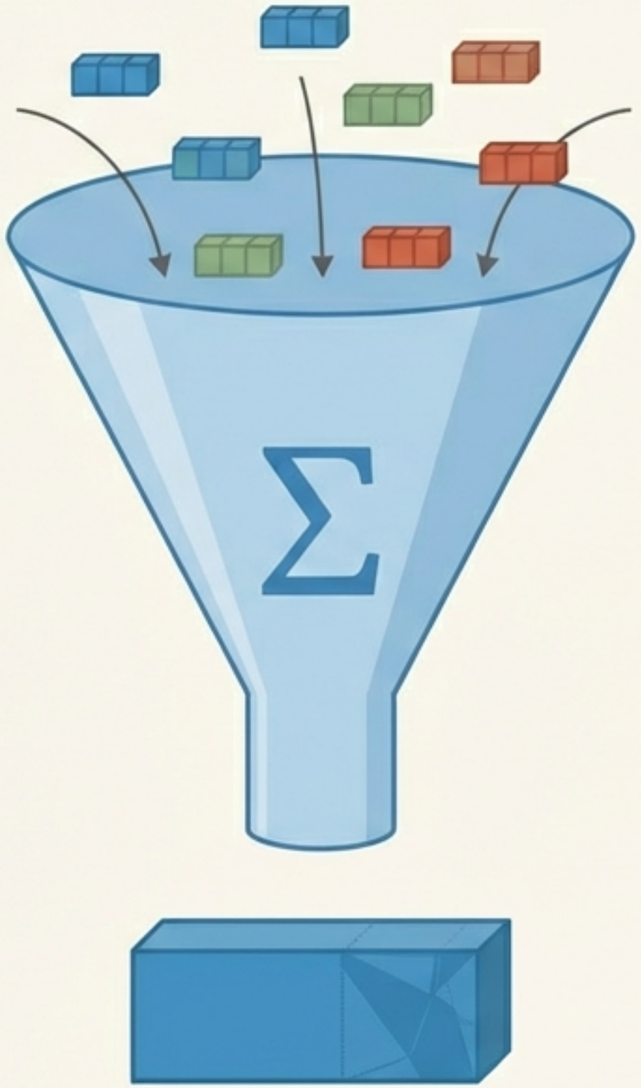
$$\mathbf{d}_{\text{TF-IDF}} = \frac{\sum \text{TF-IDF}(w, D) \cdot \mathbf{e}_{\text{w}}}{\sum \text{TF-IDF}(w, D)}$$

Down-weights stop words based on document frequency.

Strategy 3: Smooth Inverse Frequency (SIF)

$$\mathbf{d}_{\text{SIF}} = \frac{1}{|D|} \sum \frac{a}{a + P(w)} \mathbf{e}_{\text{w}}$$

Crucial final step: Removal of the first principal component (subtracting the common discourse vector) dramatically improves semantic fidelity.

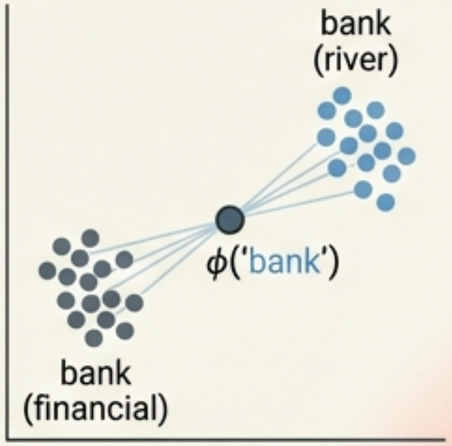


The Flaws in the Matrix

Polysemy Conflation

$$\phi(\text{'bank'}) = \sum \alpha_s \cdot \mathbf{e}^{(s)}_{\text{bank}}$$

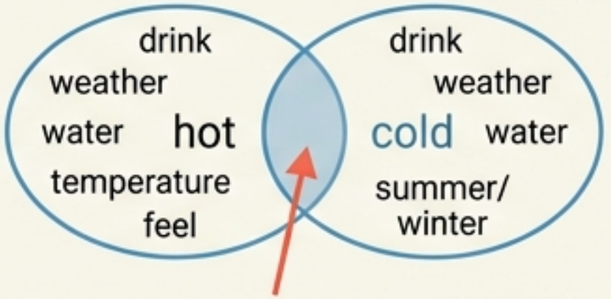
A frequency-weighted centroid that accurately represents no single sense.



The Antonym Problem

$$P(\text{context} \mid \text{hot}) \approx P(\text{context} \mid \text{cold})$$

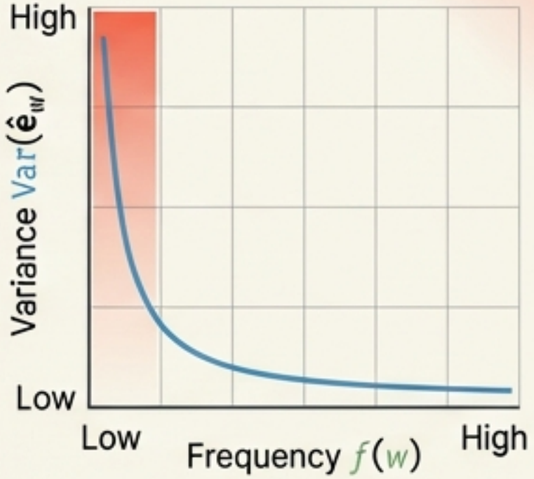
Both share context. Distributional theory **conflates opposites.**



Frequency Bias

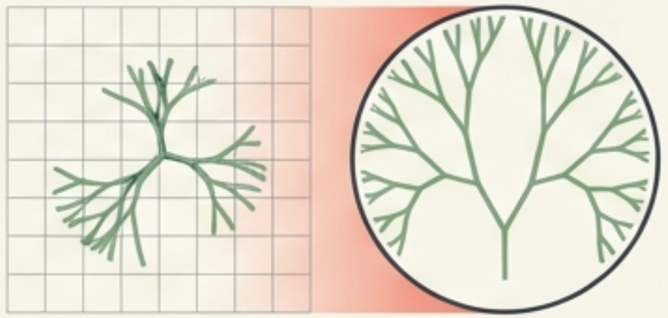
$$\text{Var}(\hat{\mathbf{e}}_w) \propto \frac{1}{f(w)}$$

Rare words yield high-variance, low-quality embeddings.



Static Geometry Constraints

Euclidean space poorly models hierarchical relationships, requiring hyperbolic (Poincaré) spaces.



The Societal Cost: Bias in Embeddings

The Mechanism of Bias: Embeddings perfectly reflect and amplify societal stereotypes present in the corpus.

Measuring Bias

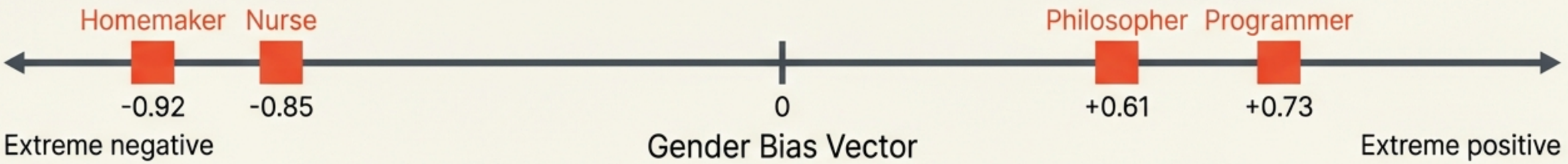
$$\text{bias}(w, b) = \cos(e_w, b)$$

For gender, $b = e_{\_he} - e_{\_she}$

Note: The WEAT (Word Embedding Association Test) provides statistical significance via permutation tests for these disparities.

Empirical Scores (Bolukbasi et al., 2016)

Homemaker: -0.92
Nurse: -0.85
Philosopher: +0.61
Programmer: +0.73



The Illusion of Debiasing

Hard Debiasing

Aims to neutralize bias by removing projection onto the bias subspace B.

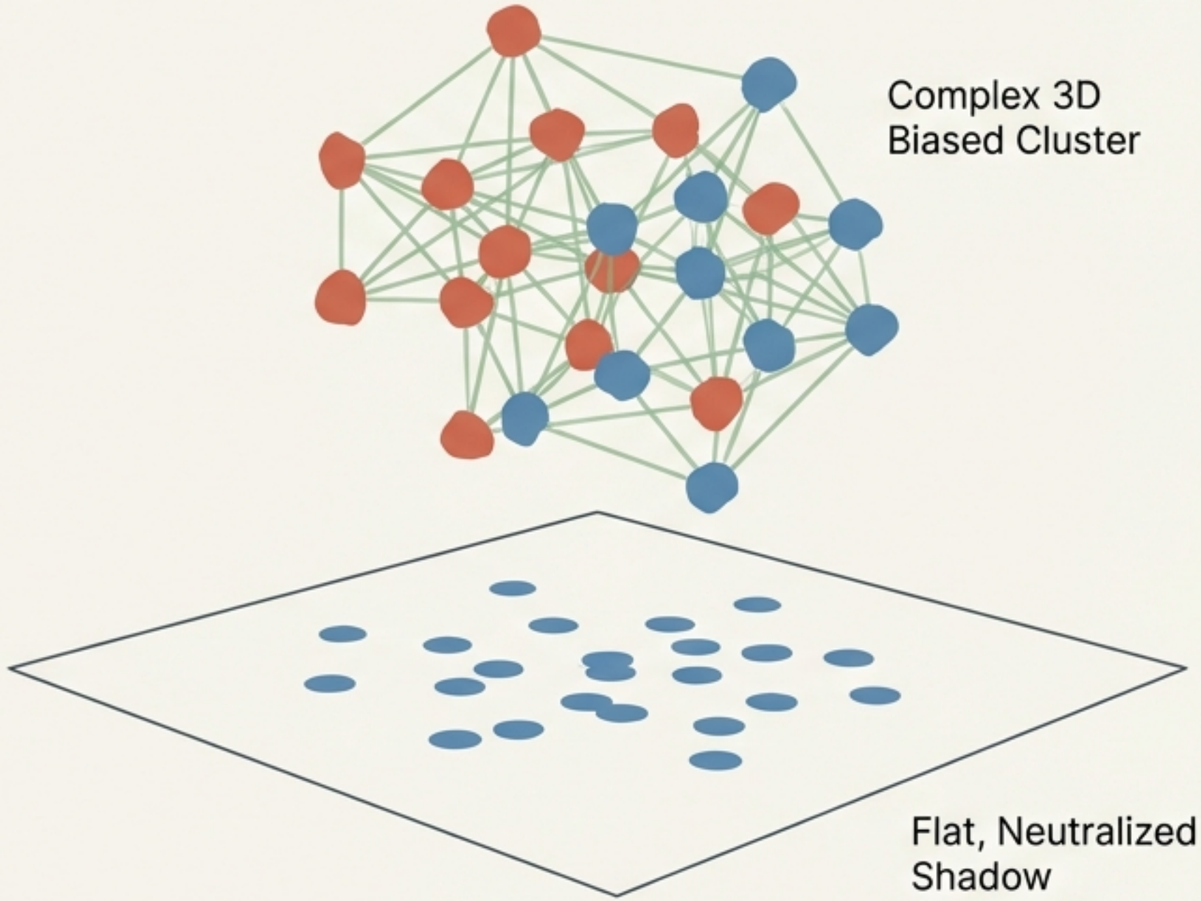
$$e_w^{debiased} = e_w - B B^T e_w$$

The Root Cause (Gonen & Goldberg, 2019)

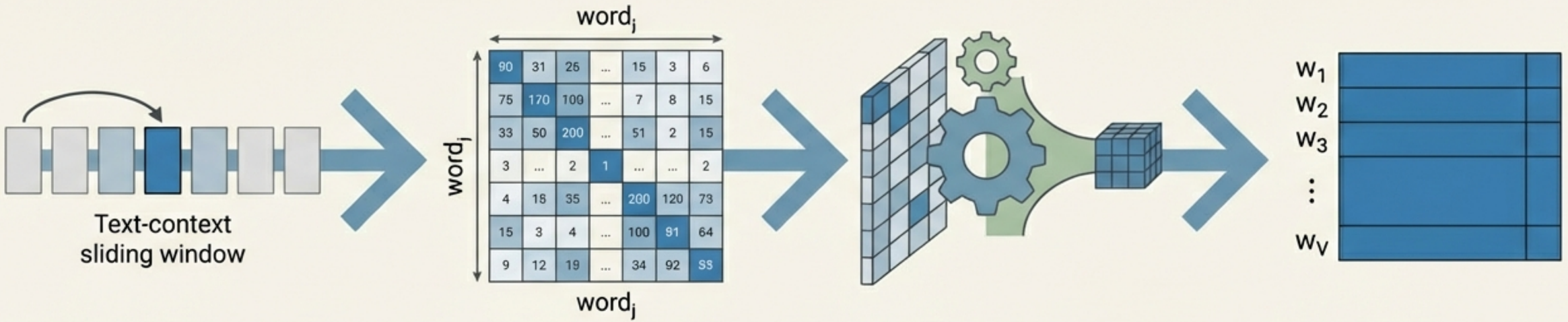
Hard debiasing is superficial. Bias is distributed across many dimensions, not just a single linear subspace.

The Symptom vs. The Disease

After linear projection, previously biased words still cluster together via nearest-neighbor structure. Projection-based debiasing hides the symptom but fails to alter the root geometry.



Unifying Perspective



1. Theory
(Distributional Hypothesis)

Meaning is context. A word is characterized by the company it keeps.

2. Data
(Co-occurrence Statistics)

Theoretical axioms operationalized into raw matrices.

3. Algorithm
(Matrix Factorization)

SGNS, GloVe, and SVD unified. All implicitly or explicitly compress the co-occurrence matrix.

4. Output
$(\mathbf{E} \in \mathbb{R}^{V \times d})$

The final geometric encoding of semantic relationships, wholly dependent on the training corpus.