Statistical Uncertainty in Word Embeddings: GloVe-V

Andrea Vallebueno[*]
Stanford University
avaimar@stanford.edu

Cassandra Handan-Nader[*]
Stanford University
slnader@stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Daniel E. Ho[†]
Stanford University
deho@stanford.edu

Abstract

Static word embeddings are ubiquitous in computational social science applications and contribute to practical decision-making in a variety of fields including law and healthcare. However, assessing the statistical uncertainty in downstream conclusions drawn from word embedding statistics has remained challenging. When using only point estimates for embeddings, researchers have no streamlined way of assessing the degree to which their model selection criteria or scientific conclusions are subject to noise due to sparsity in the underlying data used to generate the embeddings. We introduce a method to obtain approximate, easy-to-use, and scalable reconstruction error variance estimates for GloVe (Pennington et al., 2014), one of the most widely used word embedding models, using an analytical approximation to a multivariate normal model. To demonstrate the value of embeddings with variance (GloVe-V), we illustrate how our approach enables principled hypothesis testing in core word embedding tasks, such as comparing the similarity between different word pairs in vector space, assessing the performance of different models, and analyzing the relative degree of ethnic or gender bias in a corpus using different word lists.

1 Introduction

Over the past decade, vector representations of words, or "word embeddings," have become standard ways to quantify word meaning and semantic relationships due to their high performance on natural language tasks (Mikolov et al., 2013b; Pennington et al., 2014; Levy et al., 2015). Word embeddings are now ubiquitous in a wide variety of downstream computational social science applications, including charting the semantic evolution of words over time (Hamilton et al., 2016), generating simplifications of scientific terminology (Kim et al., 2016), comparing the information density of languages (Aceves and Evans, 2024), assisting in legal interpretation (Choi, 2024), and detecting societal biases in educational texts (Lucy et al., 2020), historical corpora (Garg et al., 2018; Charlesworth et al., 2022), legal documents (Matthews et al., 2022; Sevim et al., 2023), political writing (Knoche et al., 2019), and annotator judgments (Davani et al., 2023). Task performance metrics using word embeddings also factor prominently into the evaluation of more sophisticated, multimodal artificial intelligence systems, such as brain-computer interfaces (Tang et al., 2023) and adversarial text-to-image generation (Liu et al., 2023).

Though a vast amount of research has relied on a relatively narrow set of word embedding models, no unified framework has emerged for representing statistical uncertainty in how accurately the word embeddings reconstruct the relationships implied by the sample of word co-occurrences. Intuitively, we should be less certain about a word's position in vector space the less data we have on its co-occurrences in the raw text (Ethayarajh et al., 2019a,b). While it is generally standard practice in the social and natural sciences to check results for statistical significance, the vast majority of applications have relied exclusively on point estimates of embeddings, ignoring uncertainty even when training vectors over smaller corpora (e.g., Sevim et al., 2023; Knoche et al., 2019). In select applications, approaches have ranged from a bootstrap on the *documents* in the training corpus (Lucy et al., 2020), to permutations of the *word list* or lexicon (Caliskan et al., 2017; Garg et al., 2018). While useful, such bootstrap and permutation approaches are computationally intractable on large datasets. Moreover, they address uncertainty from document or lexicon selection, even though embeddings are parameters of a data generating process on word co-occurrences, not on collections of documents or sets of words (Ethayarajh et al., 2019b). Until

---
[*]Equal contribution.
[†]Corresponding author.

now, accounting for the fundamental uncertainty from the data-generating process has eluded the NLP community.

To fill this gap, we develop GloVe-V, a scalable, easy-to-use, computationally efficient method for approximating reconstruction error variances for the GloVe model (Pennington et al., 2014), one of the most widely used word embedding models. Our approach leverages the core insight that if context vectors and constant terms are held fixed at optimal values, GloVe word embeddings are the optimal parameters for a multivariate normal probability model on a weighted log transformation of the rows of the co-occurrence matrix. If we assume that the rows of the co-occurrence matrix are independent given their context vectors and constant terms, the word embedding variances according to this likelihood are computationally tractable on large vocabularies. This assumption is reasonable and is also employed in other settings, such as measuring the influence of particular documents on downstream embedding statistics (Brunet et al., 2019). Such variance estimates enable researchers to conduct rigorous assessments of model performance and principled statistical hypothesis tests on downstream tasks, responding to the need to account for statistical uncertainty and significance testing in natural language processing and machine learning (Card et al., 2020; Dror et al., 2020; Liao et al., 2021; Bowman and Dahl, 2021; Ulmer et al., 2022).

Our contributions are threefold: (a) we provide the statistical foundations for a principled notion of reconstruction error variances for GloVe word embeddings; (b) we show that incorporating uncertainty can change conclusions about textual similarities, model selection, and textual bias; (c) we provide a data release including pre-computed word embeddings and variances for the most frequently occurring words in the Corpus of American English (COHA), the largest corpus of historical American English that is widely used to track the usage and linguistic evolution of English terms over time (e.g., Ng et al., 2015; Newberry et al., 2017; Garg et al., 2018; Xiao et al., 2023; Charlesworth and Hatzenbuehler, 2024).[1]

---

[1]Our code (github.com/reglab/glove-v) and data (nlp.stanford.edu/projects/glove) repositories include data, tutorials, documentation and code to apply the GloVe-V framework to any downstream task. To make our approach more readily accessible to researchers, we release pre-computed embeddings and variances for COHA (1900-1999), intended for academic use. We also plan to make these

2   Background on GloVe

We use upper case bold letters for matrices $\mathbf{X}$, lower case bold letters for vectors $\mathbf{x}$, and regular non-bolded letters for scalars x, except when indexing into a matrix or vector (i.e., the ij^{th} entry of the matrix $\mathbf{X}$ is $\mathbf{X}_{ij}$). Sets are represented by script letters $\mathcal{X}$.

Word embedding models learn a shared vector space representation of words in a corpus. The training data are word co-occurrences in the corpus, which can be represented by a $V \times V$ co-occurrence matrix $\mathbf{X}$, where $\mathbf{X}_{ij}$ is the weighted number of times word j appears in the context of word i,[2] and V is the number of words in the vocabulary. A word embedding is a vector representation of a given word that emerges from the model.

The GloVe word embedding model (Pennington et al., 2014) learns two embeddings $\mathbf{w}_k, \mathbf{v}_k \in \mathbb{R}^D$ for each word k, by minimizing the following cost function for $\mathbf{X}_{ij} > 0$:

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(\mathbf{X}_{ij})(\mathbf{w}_i^T \mathbf{v}_j + b_i + c_j - \log \mathbf{X}_{ij})^2$$

(1)

where $f(\mathbf{X}_{ij})$ is a non-negative weighting function with properties that ensure that very rare or very frequent co-occurrences do not receive too much weight, b_i and c_j are constant terms associated with word i and j respectively.[3] The vectors $\mathbf{v}_k$ are called "context" vectors and $\mathbf{w}_k$ are called "center" vectors, representing that word co-occurrences are defined based on words that appear within a fixed context window around a center word. The original implementation computed $\mathbf{w}_k + \mathbf{v}_k$ in a post-processing step to obtain a single embedding for a word k. In this paper, we focus on the center vector $\mathbf{w}_k$ as the embedding of interest for word k.[4]

---

data products available for general use for larger corpora (e.g. Wikipedia & Gigaword).

[2]Co-occurrence terms are usually weighted by the inverse of their distance from the center word.

[3]We follow the approach of the original authors and use

$$f(x) = \begin{cases} (x/100)^{3/4} & \text{if x} < 100 \\ 1 & \text{otherwise} \end{cases}$$

[4]While summing the context and center vectors can give useful performance gains, it is not always even a win. See the extensive discussion in (Assylbekov and Takhanov, 2019; Levy et al., 2015, secs. 3.3 and 6).
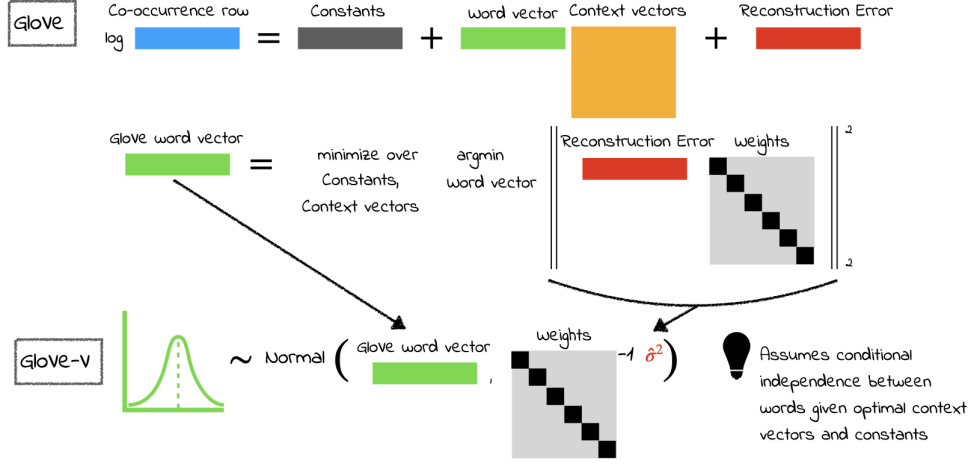
Figure 1: Conceptual diagram of the Glove-V method for one word. The top two rows illustrate the structural form and estimation of the original GloVe model (Pennington et al., 2014), which models each row of a logged, weighted co-occurrence matrix as the product of a word vector and context vectors, plus constant terms. As shown in the third row, GloVe-V creates a distribution for the optimal GloVe word vector using the reconstruction error found through the GloVe minimization procedure. These distributions can be efficiently computed word-by-word by assuming conditional independence between words given the optimal context vectors and constants.

3 Variance Derivation for GloVe-V

We now derive the GloVe-V variance estimator by recasting the optimization problem and recovering a probabilistic interpretation of GloVe embeddings.

3.1 Reformulating the GloVe optimization problem

The GloVe optimization problem using the cost in Equation 1 is an element-wise weighted low-rank approximation problem that can be broken into outer and inner minimization problems (e.g., Markovsky, 2012):

$$\min_{\substack{\mathbf{b}\in\mathbb{R}^V, \\ \mathbf{c}\in\mathbb{R}^V, \\ \mathbf{V}\in\mathbb{R}^{D\times V}}} \quad \min_{\mathbf{W}\in\mathbb{R}^{D\times V}} \|\mathbf{F}^{1/2}\odot\mathbf{R}\|_F^2$$

$$s.t. \tag{2}$$

$$\mathbf{R}_{ij} = \begin{cases} \log\mathbf{X}_{ij} - \mathbf{w}_i^T\mathbf{v}_j - b_i - c_j, & \mathbf{X}_{ij} > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\mathbf{F}_{ij} = f(\mathbf{X}_{ij})$, $\mathbf{w}_i$ and $\mathbf{v}_j$ are the ith and j^{th} columns of matrices $\mathbf{W}$ and $\mathbf{V}$ respectively, b_i and c_j are the ith and j^{th} elements of vectors $\mathbf{b}$ and $\mathbf{c}$ respectively, $\odot$ is the element-wise product, and $\|\cdot\|_F^2$ is the squared Frobenius norm.

Holding the choice of $(\mathbf{b}, \mathbf{c}, \mathbf{V})$ from the outer minimization fixed at their globally optimal values $(\mathbf{b}^*, \mathbf{c}^*, \mathbf{V}^*)$, the inner minimization to find the optimal center vectors $\mathbf{W}^*$ decomposes to V weighted least squares objectives that can be solved independently:

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{D\times V}, \\ \mathbf{b}=\mathbf{b}^* \\ \mathbf{c}=\mathbf{c}^* \\ \mathbf{V}=\mathbf{V}^*}} \|\mathbf{F}^{1/2}\odot\mathbf{R}\|_F^2 = \sum_i \min_{\substack{\mathbf{w}_i\in\mathbb{R}^D, \\ \mathbf{b}=\mathbf{b}^* \\ \mathbf{c}=\mathbf{c}^* \\ \mathbf{V}=\mathbf{V}^*}} \|\mathbf{F}^{1/2}\mathbf{r}_i\|_2^2$$

$$= \sum_i \min_{\substack{\mathbf{w}_i\in\mathbb{R}^D, \\ \mathbf{b}=\mathbf{b}^* \\ \mathbf{c}=\mathbf{c}^* \\ \mathbf{V}=\mathbf{V}^*}} \|\mathbf{F}_{\mathcal{K}_i}^{1/2}[\log\mathbf{x}_i - \mathbf{V}_{\mathcal{K}_i}^{*T}\mathbf{w}_i - b_i^*\mathbf{1} - \mathbf{c}_{\mathcal{K}_i}^*]\|_2^2 \tag{3}$$

where $\mathbf{r}_i$ is the ith row of $\mathbf{R}$, $\mathcal{K}_i = \{j : \mathbf{X}_{ij} > 0\}$, the set of column indices with non-zero co-occurrences for word i, $\mathbf{V}_{\mathcal{K}_i}^*$ is a matrix whose columns belong to the set $\{\mathbf{v}_j^* : j \in \mathcal{K}_i\}$, $\mathbf{c}_{\mathcal{K}_i}^*$ is a vector whose elements belong to the set $\{c_j^* : j \in \mathcal{K}_i\}$, $\mathbf{F}_{\mathcal{K}_i} = \text{diag}(\{\mathbf{F}_{ij} : j \in \mathcal{K}_i\})$, and $\mathbf{x}_i$ is the vector whose elements belong to the set $\{\mathbf{X}_{ij} : j \in \mathcal{K}_i\}$.

3.2 The closest natural probabilistic model

Recasting the GloVe problem as a collection of independent weighted least squares objectives allows us to take advantage of a standard probabilistic interpretation that corresponds to weighted least square models (Romano and Wolf, 2017). Equation 3 implies that, conditional on the optimal context vector subspace spanned by $\mathbf{V}^*$ and optimal constant vectors $(\mathbf{b}^*, \mathbf{c}^*)$, the optimal center vectors $\mathbf{W}^*$ are also optimal for a multivariate normal model on the non-zero elements of the rows of $\mathbf{X}$:

$$\log\mathbf{x}_i|b_i^*, \mathbf{c}_{\mathcal{K}_i}^*, \mathbf{V}_{\mathcal{K}_i}^* = b_i^*\mathbf{1} + \mathbf{c}_{\mathcal{K}_i}^* + \mathbf{V}_{\mathcal{K}_i}^{*T}\mathbf{w}_i + \mathbf{e}_i$$

$$\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{F}_{\mathcal{K}_i}^{-1}\sigma_i^2) \tag{4}$$

where we always have $(\mathbf{F}_{\mathcal{K}_i})_{kk} > 0$ for $k = 1, \ldots, |\mathcal{K}_i|$ due to the fact the $\mathbf{X}_{ij} > 0$ for $j \in \mathcal{K}_i$, and we assume that the rows of $\log \mathbf{X}$ are independent given the optimal parameters, i.e., $\log \mathbf{x}_i | \mathbf{b}^*, \mathbf{c}^*, \mathbf{V}^* \perp \log \mathbf{x}_j | \mathbf{b}^*, \mathbf{c}^*, \mathbf{V}^*$ for $i \neq j$.[5] Then, under standard assumptions for weighted least squares estimators (see Romano and Wolf, 2017), the covariance matrix for $\mathbf{W}$ simplifies into a $VD \times VD$ block diagonal matrix with the ith $D \times D$ block given by:

$$\mathbf{\Sigma}_i = \sigma_i^2 \left(\sum_{j \in \mathcal{K}_i} f(\mathbf{X}_{ij}) \mathbf{v}_j^* (\mathbf{v}_j^*)^T \right)^{-1} \quad (5)$$

We can then estimate σ_i^2, the reconstruction error for word i, with the plug-in estimator:

$$\hat{\sigma}_i^2 = \frac{1}{|\mathcal{K}_i| - D} \sum_{j \in \mathcal{K}_i} f(\mathbf{X}_{ij})(\mathbf{R}_{ij}^*)^2$$

where $\mathbf{R}_{ij}^* = \log \mathbf{X}_{ij} - \mathbf{w}_i^{*T} \mathbf{v}_j^* - b_i^* - c_j^*$ is the residual error of co-occurrence term ij using the optimal GloVe parameters. Figure 1 summarizes the full approach.

3.3 Estimation

The covariance estimator in Equation 5 is only valid for words that co-occur with a greater number of unique context words $|\mathcal{K}_i|$ than the dimensionality of the vectors D. A simple way to increase the coverage of the variances in smaller corpora is to reduce the dimensionality of the word vectors. However, when $|\mathcal{K}_i| \approx D$, numerical problems with computing the inverse in Equation 5 are likely to occur even if it is technically possible to compute an inverse. To address these numerical issues, for each word whose Hessian block $\mathbf{H}_i = \sum_{j \in \mathcal{K}_i} f(\mathbf{X}_{ij}) \mathbf{v}_j^* (\mathbf{v}_j^*)^T$ has a condition number that implies numerical error in excess of 1e-10 in its inverse, we instead compute the Moore-Penrose pseudo-inverse of $\mathbf{H}_i$ as $\mathbf{V}\mathbf{\Lambda}^+\mathbf{U}^T$ (Golub and Van Loan, 2013), where $\mathbf{H}_i = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ is the singular value decomposition of $\mathbf{H}_i$ and $\mathbf{\Lambda}_{jj}^+ = 1/\mathbf{\Lambda}_{jj}$ if $\mathbf{\Lambda}_{jj} > 1\text{e-}3 \times \max_j \mathbf{\Lambda}_{jj}$ and 0 otherwise. This technique effectively drops dimensions that are predominantly noise when computing the inverse.

3.4 Propagating uncertainty

With this derivation in hand, propagating variance to downstream tasks is straightforward. For differentiable test statistics, such as the cosine similarity between two word embeddings, the most computationally efficient approach is to use the delta method for asymptotic variances (van der Vaart, 2000). Using a first-order Taylor series approximation to the test statistic, the delta method states that if $\sqrt{n}(\mathbf{W} - \hat{\mathbf{W}})$ converges to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then $\sqrt{n}(\phi(\mathbf{W}) - \phi(\hat{\mathbf{W}}))$ converges to $\mathcal{N}(\mathbf{0}, \phi'(\mathbf{W})^T \mathbf{\Sigma} \phi'(\mathbf{W}))$, where $\phi(\cdot)$ is a differentiable function of $\mathbf{W}$ and $\phi'(\cdot)$ is its gradient with respect to $\mathbf{W}$. If the test statistic only depends on a subset of words in the vocabulary, the computation is quite efficient due to the fact that the gradient will be sparse.

For a broader class of test statistics, researchers can repeatedly draw from the estimated multivariate normal distribution in Equation 5 and recalculate the test statistic of interest (Tanner, 1996), which is much more computationally efficient than a bootstrap on the full embedding model.[6] This approach generates very similar results to those of the delta method, depending on the nature of the test statistic.[7]

4 Results

To build intuition and demonstrate the usefulness of GloVe-V variances, we now provide empirical results using the Corpus of Historical American English (COHA) for the 20th century, which contains English-language texts from a balanced set of genres (fiction, non-fiction, magazines, and newspapers) from 1900–1999 (Davies, 2012).[8]

We trained multiple GloVe models for these empirical results. First, we trained a single model on the unaltered COHA (1900-1999) corpus, which we use for all of the GloVe-V results. Second, we trained 100 GloVe models on document-level bootstrap samples of the COHA (1900-1999) corpus.[9] We trained each 300-dimensional GloVe model (132M parameters for a vocabulary size of approx-

---

[5] Brunet et al. (2019) rely on the same assumption to make their method for approximating the influence of a group of documents on word association statistics computationally tractable.

[6] Appendix A.1 provides example code for the repeated draws approach, using the cosine similarity of a word pair as an example.

[7] For example, differences can arise due to non-linear functions for which $E[f(X)] \neq f(E[X])$ (see Jensen, 1906).

[8] The data were accessed under a standard academic license in accordance with the data usage restrictions at https://www.corpusdata.org/restrictions.asp.

[9] In both cases, we pre-processed the corpus to lowercase all tokens and drop non-alphabetic characters.
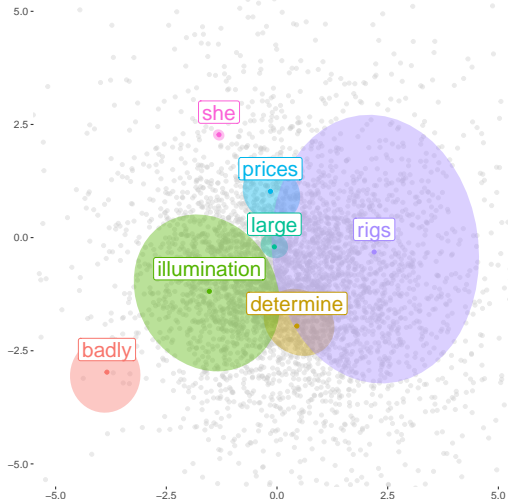
Figure 2: **Uncertainty in word embedding locations.** Two-dimensional representations of GloVe word embeddings trained on COHA (1900–1999), along with ellipses drawn around 100 draws from the estimated multivariate normal distribution from Equation 4 for a random subset of words. Lower frequency words like "rigs" and "illumination" have more uncertainty in their estimated positions in the vector space than high frequency words like "she" and "large."
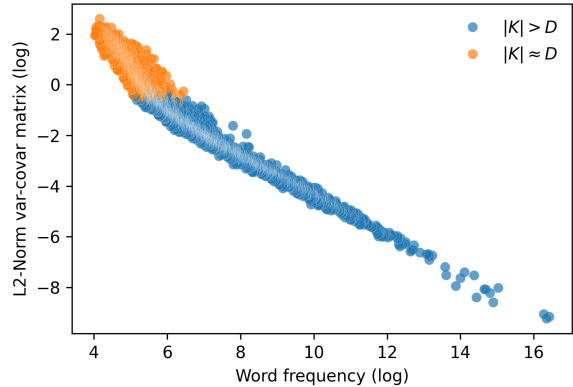

Figure 3: **Word-level relationship between GloVe-V variances and frequency on COHA (1900–1999).** L2-norm of the diagonal of $\hat{\Sigma}$ from Equation 5 (x-axis, on a $\log_{10}$ scale) plotted against logged word frequencies (y-axis, on a $\log_{10}$ scale) for a subset of 5,000 words randomly sampled in proportion to word frequency. The variances for words colored in orange are computed as discussed in Section 3.3.

imately 219,000 words) for 80 iterations, with a symmetric context window of 8 words, and using the following default hyperparameters of the official GloVe model implementation (Pennington et al., 2023): initial learning rate of 0.05, and $\alpha = 0.75$ and $x_{\max} = 100$ for the weighting function. The training of each model consumed about 40 minutes on a workstation equipped with an AMD Milan 7543 @ 2.75 GHz processor CPU, using 48 CPU cores.

The empirical examples we provide in Section 5 show how GloVe-V variances can move researchers in NLP towards emerging best practices for incorporating hypothesis testing in natural language tasks and downstream analyses (Card et al., 2020).

4.1 Uncertainty in word embedding locations

We first show that the variances can represent reconstruction error uncertainty in the locations of individual words in vector space due to data sparsity. Figure 2 plots a two-dimensional representation of the word embeddings, with ellipses drawn around 100 draws from the estimated multivariate normal distribution in Equation 4 for a random subset of words. The size of the ellipses reflects the fact that, based on the underlying co-occurrence matrix, we are more certain about the positions of higher fre-

quency words like "she" and "large" than lower frequency words like "illumination" and "rigs." The higher uncertainty for lower frequency words is a structural feature of the estimated covariance matrices themselves. Figure 3 demonstrates this feature by plotting the word-level frequency (x-axis, on a $\log_{10}$ scale) against the L2-norm of the diagonal of the estimated $\hat{\Sigma}$ from Equation 5 (y-axis, on a $\log_{10}$ scale) for a random subset of words (sampled in proportion to their frequency). The magnitude of the variances for the word embedding parameters decreases smoothly as the word frequency increases. Where $|\mathcal{K}_i| \approx D$ (highlighted in orange in Figure 3), the estimation approach described in Section 3.3 provides a reasonable estimate of the variance.

4.2 Comparison to document bootstrap

We now provide intuition for why the GloVe-V variances may in many instances be preferable to the document bootstrap for hypothesis testing on downstream test statistics. The document bootstrap is a computationally intensive approach to capturing word embedding instability that repeatedly resamples documents from the corpus, recomputes the word embeddings, and recalculates the test statistic of interest (Antoniak and Mimno, 2018). To re-purpose this approach in order to conduct a hypothesis test, we must subscribe to the uncertainty framework that the corpus itself is randomly sampled from a hypothetical population of documents.
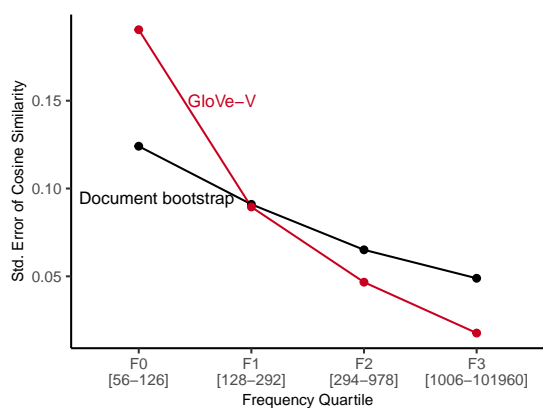
Figure 4: **Comparison between document bootstrap and GloVe-V standard errors for cosine similarity.** The average standard error of the cosine similarity between 1,600 randomly sampled word pairs (y-axis) as a function of the frequency for the word pair (x-axis with word frequency ranges in brackets), using the document bootstrap approach and GloVe-V using the delta method. The GloVe-V standard errors are more sensitive to word frequency and are more efficient to compute.

However, this framework does not match the statistical micro-foundations under which the embeddings themselves are estimated as described in Section 3. For this reason, the document bootstrap does not always capture uncertainty due to poor reconstruction of the original co-occurrences for different words, while GloVe-V is designed for this very purpose.

Figure 4 shows that document-level uncertainty can either underestimate or overestimate the variance of a downstream test statistic compared to the reconstruction error uncertainty given by GloVe-V, depending on the distribution of words across documents. A word that is used infrequently but in the same way across many documents may have low document-level uncertainty because each bootstrap sample will yield similar (but sparse) co-occurrence counts for that word, even if the reconstruction error remains high for each bootstrapped estimate. Conversely, a word that is extremely common in only a few documents may have high document-level uncertainty because many bootstrap samples will drop the majority of documents containing that word, even if the reconstruction error is low when all documents are included. Rather than choose one over the other, researchers can use each method for different purposes: tools like the document bootstrap or more computationally efficient analogs (e.g., Brunet et al., 2019) can be used to assess sensitivity of results to particular documents, while
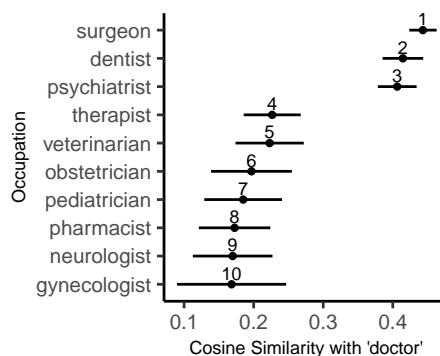

Figure 5: **Nearest neighbors with uncertainty.** Healthcare occupations (y-axis) ranked by their cosine similarity with "doctor" (x-axis), with the nearest neighbor ranking based on the point estimate above each point, and 95% GloVe-V uncertainty intervals.

GloVe-V can be used to conduct hypothesis tests under a coherent statistical framework, holding the corpus fixed.

5 GloVe-V enables principled significance testing

We now show how GloVe-V enables statistical significance testing, addressing the increasing recognition for NLP to move beyond point estimates alone (Card et al., 2020; Liao et al., 2021). GloVe-V can also help researchers assess when a corpus is underpowered for specific inferences, as we illustrate below.

5.1 Uncertainty in k nearest neighbors

Word similarity, including k nearest neighbor lists, informs performance evaluation for both embedding models and more sophisticated artificial intelligence systems (e.g., Mikolov et al., 2013a; Levy and Goldberg, 2014a; Linzen, 2016; Borah et al., 2021; Tang et al., 2023; Liu et al., 2023). Using only point estimates to evaluate word similarity, however, leaves the researcher with no sense of which word similarities are inherently less certain because they are based on less co-occurrence data in the underlying corpus. Uncertainty in neighbor rankings is particularly consequential for word similarity tasks, which depend on the ranking of different word pairs, and for word analogy tasks, which are typically solved by finding nearest neighbors in the embedding space.[10] As an example of

---

[10]In the case of analogy tasks, the solution typically relies on the nearest neighbor to some linear transformation of the word embeddings belonging to the words in the analogy (Levy and Goldberg, 2014a), rather than the nearest neighbor to a
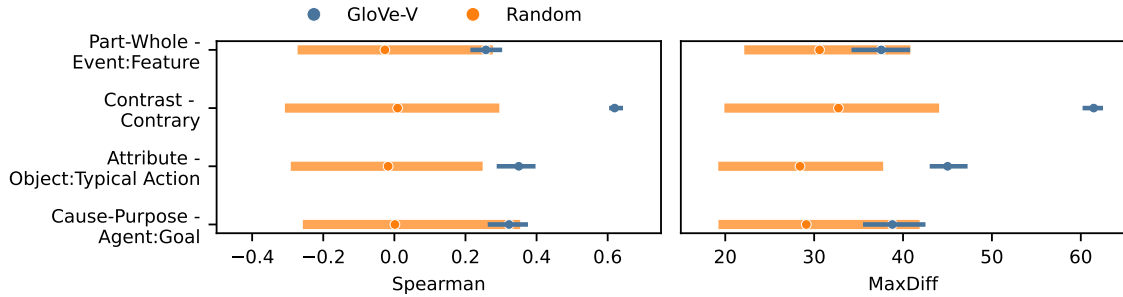
Figure 6: **Accounting for uncertainty in word embedding performance assessments using the SemEval-2012 Task 2 of Jurgens et al. (2012) on COHA 1900–1999.** The task measures the degree of relational similarity of word pairs using the Spearman correlation and the MaxDiff choice procedure on a taxonomy that comprises 79 types of relations across 10 different classes (*e.g.*, contrast, part-whole, cause-purpose). We present GloVe-V results on a subset of relations, along with a Random baseline that randomly rates the word pairs in each relation.

this dilemma, Figure 5 plots the cosine similarity between "doctor" and a list of healthcare occupation words along with GloVe-V uncertainty intervals. Based on the point estimates, we can assign a nearest-neighbor rank to each word by their proximity to "doctor" (printed above the point estimate in Figure 5). However, for the top three neighbors (and between neighbours 4 through 10), we cannot statistically distinguish the ranks (e.g., $p = 0.10$ for the difference in the cosine similarity of "doctor" and "surgeon" relative to that of "doctor" and "dentist"). Which neighbor is the "nearest" is therefore subject to considerable uncertainty that would be invisible without incorporation of the GloVe-V variances.

5.2 Uncertainty in model performance

Performance on analogy tasks is a canonical approach to word embedding model evaluation (Mikolov et al., 2013b; Pennington et al., 2014; Levy et al., 2015). Closed-list relational similarity tasks such as SemEval-2012 (Jurgens et al., 2012) are structured so that models can be benchmarked against random baselines, in which pairs of words are randomly related to each other to establish a lower bound on expected performance. Figure 6 presents the performance of GloVe compared to a random benchmark using two evaluation metrics on four relational similarity tasks (see Jurgens et al., 2012, for details on the tasks and metrics). While the point estimates for performance suggest that GloVe outperforms the random baseline on all tasks, adding uncertainty to both point estimates reveals that we can only claim significantly higher performance than random on two of the four rela-

specific word as illustrated in our example.

tions ($p = 0.08$, $p < 0.001$, $p = 0.02$, $p = 0.08$ for relations 1–4, respectively, for the MaxDiff metric). The GloVe-V intervals also allow us to distinguish the performance of GloVe across different relational similarity tasks. While we can say that GloVe performs better on "Contrast" vs. "Cause-Purpose" ($p = 0.03$ for the MaxDiff metric), we cannot claim that it does better on "Cause-Purpose" than "Part-Whole" ($p = 0.92$), even though the point estimates suggest better performance on the former task.[11]

5.3 Uncertainty in word embedding bias

Measurement of societal biases in text is a popular downstream application using word embeddings (e.g., Garg et al., 2018; Lucy et al., 2020; Charlesworth et al., 2022; Matthews et al., 2022; Sevim et al., 2023). These types of studies compare similarity between curated sets of words to test the prevalence of societal biases in text. For example, if a set of female-oriented words is closer to a set of family-oriented words than to career-oriented words, and this relationship is stronger relative to the same comparison using male-oriented words, that is evidence of a gender bias (Bolukbasi et al., 2016; Caliskan et al., 2017). To represent uncertainty in these comparisons, researchers typically use a permutation test or bootstrap on the words included in each set (e.g., Caliskan et al., 2017;

---

[11]The relations contain about about 27 to 45 word pairs: 1) *Contrast - Contrary*: pairs of opposite words (e.g., "dull" and "bright"); 2) *Part-Whole - Event: Feature*: pairs in which one word is a part of the event given by the first word (e.g., "carnival" and "rides"); 3) *Attribute - Object: Typical Action*: pairs in which one word is a characteristic action of the other (e.g., "heart" and "beat"); 4) *Cause-Purpose - Agent: Goal*: pairs in which one word is a typical objective of the agent given by the first word (e.g., "painter" and "portrait").
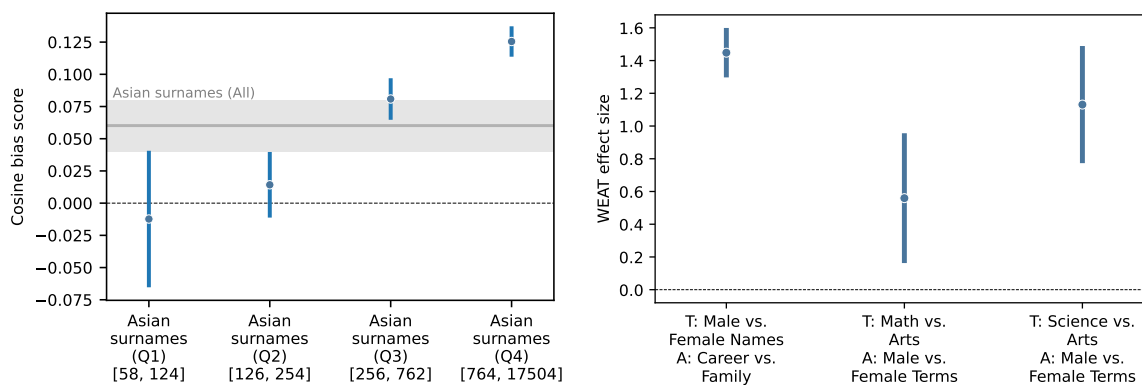
Figure 7: **Ethnicity and gender bias scores. a)** Average Asian bias scores with GloVe-V uncertainty using the cosine bias score of Garg et al. (2018) in COHA 1900–1999 for different Asian surname lists. The gray line and shaded gray area represent the point estimate and 95% GloVe-V uncertainty interval, respectively, of the bias score on a master list of Asian surnames. The points and error bars in blue represent the bias score computed on different subsets of the list, grouped according to the number of times they appear in the corpus. **b)** Gender bias scores for three types of bias tests using the WEAT effect size of Caliskan et al. (2017) with GloVe-V uncertainty.

Garg et al., 2018), but others have noted that these types of uncertainty measures (which account for uncertainty in word selection) are not designed to account for the sparsity of word co-occurrences that form the basis for the comparisons (Ethayarajh et al., 2019b).

It is especially important to account for uncertainty due to sparsity in applications where the analysis relies on infrequently occurring words such as surnames, which are often used to measure demographic bias (e.g., Caliskan et al., 2017; Garg et al., 2018; Swinger et al., 2019). Researchers typically drop lower frequency surnames altogether from their analyses (e.g., Garg et al., 2018) because they have no way of representing the higher uncertainty in the embedded positions for lower frequency surnames using only point estimates. But such curation runs the risk of sacrificing the representativeness of the word lists involved and the generalizability of the conclusions (Antoniak and Mimno, 2021). Using a measure of anti-Asian bias in the COHA corpus based on Garg et al. (2018), the left panel of Figure 7 shows how GloVe-V variances can automatically provide information on co-occurrence sparsity for researchers.[12] The bias measure on the y-axis computes the average cosine similarity between a set of Asian surnames and a set of 20 Otherization words,[13] relative to a set of White surnames. A more positive bias score indi-

cates that Asian surnames are more closely related to these negative Otherization words compared to White surnames.

Figure 7 shows that the anti-Asian bias estimate in COHA becomes more positive for more frequently appearing surnames, such that relying only on the most frequent surnames produces an exaggerated result relative to the full set of Asian surnames.[14] This is likely due to the fact that the most frequently occurring surnames tend to be from historical figures such as "Ghandi," "Mao," and "Mohammed," which are clearly not representative of the entire class of Asian surnames. Using GloVe-V, low and high frequency words can be seamlessly combined into a single bias interval that represents the combined uncertainty in all the estimated word positions (shown as a gray interval on the plot), without having to drop any surnames and sacrifice generalizability.[15]

GloVe-V intervals can also be useful for studying high frequency word lists because they allow researchers to make statistical comparisons between *types* of bias. The right panel of Figure 7 provides an example of three gender bias queries with GloVe-V intervals for the Word Embedding Association Test (WEAT) effect size, a cosine-similarity-

---

[12]Appendix A.2 provides delta method derivations for the bias statistics used in Figure 7.

[13]This word list is primarily composed of adjectives used to describe people as outsiders, such as *monstruous*, *devious*, and *bizarre*.

[14]See Appendix A.3 for details on how we compiled this surname list.

[15]GloVe-V intervals can also help researchers determine when a bias time trend can be supported by the co-occurrence data. Estimating the anti-Asian bias separately for each decade of COHA with GloVe-V uncertainty intervals, for example, reveals no significant trend in the bias score over time, suggesting that time trends such as those reported in Garg et al. (2018) should also be checked for statistical significance.

based test (Caliskan et al., 2017): (a) male vs. female names and words related to *career* vs. *family*; (b) male vs. female terms and words related to *math* vs. *arts*; and (c) male vs. female terms and words related to *science* vs. *arts*. While the point estimate for type (a) is higher than those for both (b) and (c), the GloVe-V intervals allow us to reject the null that (a) is equal to (b) ($p < 0.001$), but not the null that (a) is equal to (c) ($p = 0.11$). In this case, the GloVe-V intervals guard against making unsubstantiated claims about which types of bias are strongest.

6 Conclusion

In this paper, we have derived, computed, and demonstrated the utility of GloVe-V, a new method to represent uncertainty in word embedding locations using the seminal GloVe model. The GloVe-V variances provide researchers with the ability to easily propagate uncertainty in the word embeddings to downstream test statistics of interest, such as word similarity metrics that are used in both evaluation and analyses involving word embeddings. Unlike methods such as the document and word list bootstrap, the method is computationally efficient even on large corpora and represents uncertainty due to sparsity in the underlying co-occurrence matrix, which is often invisible in downstream analyses that use only the estimated word embeddings.

As we have shown in Section 5, incorporating uncertainty into downstream analyses can have consequential impacts on the conclusions researchers draw, and should be a best practice moving forward for studies that use word embeddings to infer semantic meaning. Finally, while outside the scope of the current study, we note that the contextual word and passage representations of transformer large language models are also point estimates and similar questions of embedding uncertainty apply when using such models as well.

Limitations

While useful in many applications, the GloVe-V method comes with certain limitations. First, the variances can only be computed for words whose number of context words exceeds the embedding dimensionality. This limitation can easily be minimized by reducing the dimensionality of the vectors for small corpora; for example, variances can be computed for 96% of the word embeddings in the relatively small *New York Times* Annotated Cor-

pus (NYT) with 50-dimensional vectors, compared to 36% of embeddings with 300-dimensional vectors. Second, researchers need access to the co-occurrence matrix if they wish to compute the variances themselves, since it relies on an empirical estimate of the reconstruction error. Third, the methodology in this paper applies solely to the GloVe embedding model because of its statistical foundations. That said, this model is one of the most-cited word embedding models in current use and has been shown to have better stability and more intuitive geometric properties than competing models (Mimno and Thompson, 2017; Wendlandt et al., 2018).

Finally, the uncertainty captured by GloVe-V intervals maps onto the extent to which the embeddings are able to reconstruct the underlying co-occurrence matrix, which is only one of many types of uncertainty one could consider in embedded locations for words. Other types of uncertainty that are held fixed in GloVe-V include instability due to the documents included in the corpus (e.g., Antoniak and Mimno, 2018), uncertainty due to the hyper-parameters of the model (e.g., Borah et al., 2021), and statistical uncertainty in the estimated context vector positions and bias terms, which are treated as constants in the variance computation for computational tractability. Along with the conditional independence assumption on words, treating these terms as constants is necessary to reduce the number of free parameters in the model and allow a tractable variance computation. Holding free parameters fixed may underestimate the total variance in the GloVe optimization procedure. However, these sorts of independence assumptions are becoming standard practice to enable computational efficiency for models with a large number of parameters – the same assumption, for example, has been successfully employed to develop a computationally efficient approximation to a document bootstrap (Brunet et al., 2019).

Despite these limitations, we found by trying a number of other approaches (detailed in Appendix A.4) that GloVe-V strikes a desirable balance between maintaining the model's probabilistic foundations for enhanced statistical rigor, and preserving computational tractability for practical purposes.

References

Pedro Aceves and James A Evans. 2024. Human languages with greater information density have higher communication speed but lower conversation breadth. *Nature Human Behaviour*, 8:1–13.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904.

Zhenisbek Assylbekov and Rustem Takhanov. 2019. Context vectors are reflections of word vectors in half the dimensions. *Journal of Artificial Intelligence Research*, 66:225–242.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Jonathan W. Book. 2020. Training bayesian neural networks: A study of improvements to training algorithms. Master thesis.

Angana Borah, Manash Pratim Barman, and Amit Awekar. 2021. Are word embedding methods stable and should we care about it? In *Proceedings of the 32nd ACM Conference on Hypertext and social media*, pages 45–55.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4843–4855.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

Tessa ES Charlesworth and Mark L Hatzenbuehler. 2024. Mechanisms upholding the persistence of stigma across 100 years of historical text. *Scientific Reports*, 14(1):11069.

Jonathan H Choi. 2024. Measuring clarity in legal text. *University of Chicago Law Review*, 91:1.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. *Statistical Significance Testing for Natural Language Processing*. Number 45 in Synthesis Lectures on Human Language Technologies. Springer Nature.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gene H Golub and Charles F Van Loan. 2013. *Matrix Computations*. JHU press.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Rujun Han, Michael Gill, Arthur Spirling, and Kyunghyun Cho. 2018. Conditional word embedding and hypothesis testing via Bayes-by-backprop. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4890–4895. Association for Computational Linguistics.

J. L. W. V. Jensen. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(0):175–193.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364. Association for Computational Linguistics.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. Simplescience: Lexical simplification of scientific terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071.

Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2019. Identifying biases in politically biased wikis through word embeddings. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 253–257. Association for Computing Machinery. Event-place: Hof, Germany.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? A meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18. Association for Computational Linguistics.

Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. 2023. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*, 6(3):2332858420940312.

Ivan Markovsky. 2012. *Low rank approximation: Algorithms, implementation, applications*, volume 906 of *Communications and Control Engineering*. Springer.

Sean Matthews, John Hudzina, and Dawn Sepehr. 2022. Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12026–12033.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. Arxiv:1301.3781, version Number: 3.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

National Archives at San Francisco. Case files by birth country.

Mitchell G Newberry, Christopher A Ahern, Robin Clark, and Joshua B Plotkin. 2017. Detecting evolutionary forces in language change. *Nature*, 551(7679):223–226.

Reuben Ng, Heather G Allore, Mark Trentalange, Joan K Monin, and Becca R Levy. 2015. Increasing negativity of age stereotypes across 200 years:

Evidence from a database of 400 million words. *PloS one*, 10(2):e0117086.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2023. GloVe: Global vectors for word representation. Github repository: Commit a6f2b94.

Joseph P Romano and Michael Wolf. 2017. Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19.

Nurullah Sevim, Furkan Şahinuç, and Aykut Koç. 2023. Gender bias in legal corpora and debiasing it. *Natural Language Engineering*, 29(2):449–482.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.

Martin A. Tanner. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3 edition. Springer Series in Statistics. Springer New York.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aad W. van der Vaart. 2000. *Asymptotic statistics*, volume 3. Cambridge University Press.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Xiao, Naomi Baes, Ekaterina Vylomova, and Nick Haslam. 2023. Have the concepts of 'anxiety' and 'depression' been normalized or pathologized? A corpus study of historical semantic change. *PLOS one*, 18(6):e0288027.

A Appendix

A.1 Code Demonstration: Using GloVe-V for Word Pair Cosine Similarity

In this section, we illustrate how to leverage our GloVe-V variances to easily and quickly compute the distribution of a downstream test statistic. For our demonstrative example, we calculate the mean and standard deviation of the cosine similarity of a pair of words using the Method of Composition (MOC) (Tanner, 1996) or repeated draws approach discussed in Section 3.4. This represents the same downstream task as the nearest neighbor with uncertainty analysis presented in Figure 5 and discussed in Section 5.1.

As noted in the code, in this demonstration we randomly generate the GloVe embeddings and the GloVe-V variances for simplicity of exposition. However, we make available pre-computed variances for various corpora (including COHA 1900-1999) and software for users to easily download and interact with these data products in our code (`github.com/reglab/glove-v`) and data (`nlp.stanford.edu/projects/glove`) repositories.

```python
import numpy as np

# Parameters
# * K: Number of samples to draw
K = 100_000
# * d: Embedding dimensionality
d = 300

# Helper functions
def cosine_similarity(v1, v2):
    """
    Computes the cosine similarity for a
        pair of vectors.
    """
    return np.dot(v1, v2) / (np.linalg.
        norm(v1) * np.linalg.norm(v2))

def random_psd_mat(n):
    """
    Generates a random positive semi-
        definite matrix.
    """
    A = np.random.randn(n, n)
    return A.T @ A

def sample_vector(vector, variance, n):
    """
    Returns a matrix of n samples for a
        word from the Multivariate
        Normal distribution given by the
        vector and covariance matrix.
    """
    v_sam = np.random.
        multivariate_normal(mean=vector,
        cov=variance, size=n)
    return v_sam
```

```
# For simplicity, in this example we
    randomly generate the embeddings and
    GloVe-V variances. In our code  (
    github.com/reglab/glove-v) and data
    (nlp.stanford.edu/projects/glove)
    repositories, we provide code to
    download pre-computed variances for
    several corpora.
np.random.seed(42)
vectors = {
    'word1': np.random.randn(d, ),
    'word2': np.random.randn(d, )
}
variances = {
    'word1': random_psd_mat(d) * 1e-4,
    'word2': random_psd_mat(d) * 1e-4
}

# Draw K multivariate normal samples for
    each word
sample_matrix_w1 = sample_vector(
    vector=vectors['word1'],
    variance=variances['word1'],
    n=K
)
sample_matrix_w2 = sample_vector(
    vector=vectors['word2'],
    variance=variances['word2'],
    n=K
)

# Method of composition: compute the
    statistic of interest using each
    sample
moc_cs = np.sum(sample_matrix_w1 *
    sample_matrix_w2, axis=1)
moc_cs /= (np.linalg.norm(
    sample_matrix_w1, axis=1) * np.
    linalg.norm(sample_matrix_w2, axis
    =1))

# Compute mean and standard deviation of
     the statistic
print(f"Statistic on the trained
    embeddings: {round(cosine_similarity
    (vectors['word1'], vectors['word2'])
    , 3)}")
print(f"MOC Mean: {round(np.mean(moc_cs)
    , 3)}")
print(f"MOC Standard deviation: {round(
    np.sqrt(np.var(moc_cs)), 3)}")
```

Output:

```
Statistic on the trained embeddings: -0.04
MOC Mean: -0.039
MOC Standard deviation: 0.014
```

A.2 Variance Estimator Derivation for Bias Measures Using the Delta Method

A.2.1 Cosine Similarity Bias

The cosine similarity bias function of Garg et al. (2018) is the following:

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{M}_A, \mathbf{M}_W) = \frac{1}{K} \sum_i \cos(\mathbf{v}_i, \mathbf{M}_A) - \frac{1}{K} \sum_i \cos(\mathbf{v}_i, \mathbf{M}_W)$$

where $\mathbf{v}_i$ is the word vector for otherization word i, $\mathbf{M}_A$ is the mean word vector over all Asian surnames (using pre-normalized vectors), and $\mathbf{M}_W$ is the mean word vector over all White surnames (using pre-normalized vectors). There are K otherization words.

In general, the partial derivative of the cosine similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$ with respect to $\mathbf{a}$ is:

$$\frac{\partial \cos(\mathbf{a}, \mathbf{b})}{\partial \mathbf{a}} = \frac{\partial \mathbf{a}^T \mathbf{b} / \partial \mathbf{a}}{\|\mathbf{b}\| \|\mathbf{a}\|} - \frac{\cos(\mathbf{a}, \mathbf{b})}{\|\mathbf{a}\|} \cdot \partial \|\mathbf{a}\| / \partial \mathbf{a}$$

If we take $\mathbf{a}$ to be an otherization word $\mathbf{v}_i$, this gives us:

$$\frac{\partial \cos(\mathbf{v}_i, \mathbf{M}_A)}{\partial \mathbf{v}_i} = \frac{\mathbf{M}_A}{\|\mathbf{M}_A\| \|\mathbf{v}_i\|} - \cos(\mathbf{v}_i, \mathbf{M}_A) \cdot \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|^2}$$

If we take $\mathbf{a}$ to be an Asian surname vector $\mathbf{a}_j$, where $\mathbf{M}_A = \frac{1}{m} \sum_j \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|}$, this gives us:

$$\frac{\partial \cos(\mathbf{v}_i, \mathbf{M}_A)}{\partial \mathbf{a}_j} = \frac{1}{m} \left(\frac{I}{\|\mathbf{a}_j\|} - \frac{\mathbf{a}_j \mathbf{a}_j^T}{\|\mathbf{a}_j\|^3} \right)$$
$$\left[\frac{\mathbf{v}_i}{\|\mathbf{v}_i\| \|\mathbf{M}_A\|} - \cos(\mathbf{v}_i, \mathbf{M}_A) \cdot \frac{\mathbf{M}_A}{\|\mathbf{M}_A\|^2} \right]$$
$$= \frac{1}{m \|\mathbf{a}_j\| \|\mathbf{M}_A\|} \mathbf{X}_{a_j}^T (\tilde{\mathbf{v}}_\mathbf{i} - \cos(\mathbf{v}_i, \mathbf{M}_A) \cdot \tilde{\mathbf{M}}_\mathbf{A})$$

where $\mathbf{X}_{a_j} = \mathbf{I} - \tilde{\mathbf{a}}_j \tilde{\mathbf{a}}_j^T$.

So the partial derivatives with respect to each type of vector are the following:

9044

$$\frac{\partial f}{\partial \mathbf{v}_i} = \frac{1}{K\|\mathbf{v}_i\|}$$
$$\left([\tilde{\mathbf{M}}_A - \tilde{\mathbf{M}}_W] - [\cos(\mathbf{v}_i, \mathbf{M}_A) - \cos(\mathbf{v}_i, \mathbf{M}_W)] \cdot \tilde{\mathbf{v}}_i\right) := \mathbf{d}_v$$

$$\frac{\partial f}{\partial \mathbf{a}_j} = \frac{1}{Km\|\mathbf{M}_A\|}\left(\frac{I}{\|\mathbf{a}_j\|} - \frac{\mathbf{a}_j\mathbf{a}_j^T}{\|\mathbf{a}_j\|^3}\right)$$
$$\sum_i \left(\tilde{\mathbf{v}}_i - \cos(\mathbf{v}_i, \mathbf{M}_A) \cdot \tilde{\mathbf{M}}_A\right)$$
$$= \frac{1}{Km\|\mathbf{a}_j\|\|\mathbf{M}_A\|}\mathbf{X}_{a_j}^T$$
$$\sum_i \left(\tilde{\mathbf{v}}_i - \cos(\mathbf{v}_i, \mathbf{M}_A) \cdot \tilde{\mathbf{M}}_A\right) := \mathbf{d}_a$$

$$\frac{\partial f}{\partial \mathbf{w}_j} = -\frac{1}{Kn\|\mathbf{M}_W\|}\left(\frac{I}{\|\mathbf{w}_j\|} - \frac{\mathbf{w}_j\mathbf{w}_j^T}{\|\mathbf{w}_j\|^3}\right)$$
$$\sum_i \left(\tilde{\mathbf{v}}_i - \cos(\mathbf{v}_i, \mathbf{M}_W) \cdot \tilde{\mathbf{M}}_W\right)$$
$$= -\frac{1}{Kn\|\mathbf{w}_j\|\|\mathbf{M}_W\|}\mathbf{X}_{w_j}^T$$
$$\sum_i \left(\tilde{\mathbf{v}}_i - \cos(\mathbf{v}_i, \mathbf{M}_W) \cdot \tilde{\mathbf{M}}_W\right) := \mathbf{d}_w$$

where $\tilde{\mathbf{a}}$ is the normalized version of vector $\mathbf{a}$.

A.2.2 WEAT effect size

For two sets of attributes, $\mathbf{V}$ and $\mathbf{Z}$, of equal size (k) the WEAT effect size of Caliskan et al. (2017) is the following:

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{z}_1, \ldots, \mathbf{z}_k, \mathbf{a}_1, \ldots, \mathbf{a}_A, \mathbf{w}_1, \ldots, \mathbf{w}_W)$$
$$= \frac{\frac{1}{|V|}\sum_i s(\mathbf{v}_i, \mathbf{W}, \mathbf{A}) - \frac{1}{|Z|}\sum_i s(\mathbf{z}_i, \mathbf{W}, \mathbf{A})}{\text{std. dev}_{x\in V\cup Z} s(x, W, A)} := \frac{H}{G}$$

where

$$H = \sum_i \left(\frac{1}{W}\sum_j \cos(v_i, w_j) - \frac{1}{A}\sum_j \cos(v_i, a_j)\right)$$
$$- \sum_i \left(\frac{1}{W}\sum_j \cos(z_i, w_j) - \frac{1}{A}\sum_j \cos(z_i, a_j)\right)$$
$$= \sum_{v\in V} s(v, W, A) - \sum_{z\in Z} s(z, W, A)$$

and

$$G^2 = \frac{1}{|V\cup Z| - 1}$$
$$\sum_{x\in V\cup Z} \left(s(x, W, A) - \frac{1}{|Z\cup V|}\sum_y s(y, W, A)\right)^2$$
$$:= \frac{1}{|V\cup Z| - 1}\sum_{x\in V\cup Z}(s(x, W, A) - E)^2$$

Let $c'_{\mathbf{a}}(\mathbf{a}, \mathbf{b}) = \frac{\partial \cos(\mathbf{a}, \mathbf{b})}{\partial \mathbf{a}}$. We first define the following derivatives:

$$\frac{\partial s(\mathbf{v}_i, W, A)}{\partial \mathbf{v}_i} = \frac{1}{W}\sum_j c'_{\mathbf{v}_i}(\mathbf{v}_i, \mathbf{w}_j) - \frac{1}{A}\sum_j c'_{\mathbf{v}_i}(\mathbf{v}_i, \mathbf{a}_j)$$
$$\frac{\partial s(\mathbf{z}_i, W, A)}{\partial \mathbf{z}_i} = \frac{1}{W}\sum_j c'_{\mathbf{z}_i}(\mathbf{z}_i, \mathbf{w}_j) - \frac{1}{A}\sum_j c'_{\mathbf{z}_i}(\mathbf{z}_i, \mathbf{a}_j)$$
$$\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{w}_j} = \frac{1}{W}c'_{\mathbf{w}_j}(\mathbf{w}_j, \mathbf{x}), \text{ for } \mathbf{x} \in \{\mathbf{v}_i, \mathbf{z}_i\}$$
$$\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{a}_j} = \frac{-1}{A}c'_{\mathbf{a}_j}(\mathbf{a}_j, \mathbf{x}), \text{ for } \mathbf{x} \in \{\mathbf{v}_i, \mathbf{z}_i\}$$

The partial derivatives with respect to each type of vector are the following:

$$\frac{\partial f}{\partial \mathbf{v}_i} = \left(\frac{1}{G|V|} - \frac{H}{G^{\frac{5}{2}}} \cdot \frac{(s(\mathbf{v}_i, W, A) - E)}{|V\cup Z|}\right)\frac{\partial s(\mathbf{v}_i, W, A)}{\partial \mathbf{v}_i}$$
$$\frac{\partial f}{\partial \mathbf{z}_i} = \left(\frac{-1}{G|Z|} - \frac{H}{G^{\frac{5}{2}}} \cdot \frac{(s(\mathbf{z}_i, W, A) - E)}{|V\cup Z|}\right)\frac{\partial s(\mathbf{z}_i, W, A)}{\partial \mathbf{z}_i}$$
$$\frac{\partial f}{\partial \mathbf{w}_j} = \frac{1}{G}\cdot\left(\frac{1}{|V|}\sum_v \frac{\partial s(\mathbf{v}_i, W, A)}{\partial \mathbf{w}_j} - \frac{1}{|Z|}\sum_z \frac{\partial s(\mathbf{z}_i, W, A)}{\partial \mathbf{w}_j}\right)$$
$$- \frac{H}{G^{\frac{5}{2}}}\frac{1}{|V\cup Z| - 1}\sum_{x\in V\cup Z}\left((s(\mathbf{x}, W, A) - E)\cdot\left(\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{w}_j} - \frac{\partial E}{\partial \mathbf{w}_j}\right)\right)$$
$$\frac{\partial f}{\partial \mathbf{a}_j} = \frac{1}{G}\cdot\left(\frac{1}{|V|}\sum_v \frac{\partial s(\mathbf{v}_i, W, A)}{\partial \mathbf{a}_j} - \frac{1}{|Z|}\sum_z \frac{\partial s(\mathbf{z}_i, W, A)}{\partial \mathbf{a}_j}\right)$$
$$- \frac{H}{G^{\frac{5}{2}}}\frac{1}{|V\cup Z| - 1}\sum_{x\in V\cup Z}\left((s(\mathbf{x}, W, A) - E)\cdot\left(\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{a}_j} - \frac{\partial E}{\partial \mathbf{a}_j}\right)\right)$$

where

$$\frac{\partial E}{\partial \mathbf{w}_j} = \frac{1}{|V\cup Z|}\sum_{x\in V\cup Z}\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{w}_j}$$
$$\frac{\partial E}{\partial \mathbf{a}_j} = \frac{1}{|V\cup Z|}\sum_{x\in V\cup Z}\frac{\partial s(\mathbf{x}, W, A)}{\partial \mathbf{a}_j}$$

A.2.3 Delta Method

Using the derivatives computed in the sections above, the variance of the bias calculation is:

$$\text{var}(h) = \sum_i (\mathbf{d}_t)_i^T \boldsymbol{\Sigma}_i (\mathbf{d}_t)_i$$

where $h \in [f, g]$ is the bias function, t is the type of word i (e.g., $t = a$ for Asian surnames, $t = w$ for White surnames, $t = v$ for Otherization words in the case of the cosine similarity metric), and $\boldsymbol{\Sigma}_i$ is the variance-covariance matrix for the parameters of word i (Equation 5).

A.3 Asian Surname List Generation

To explore the behavior of anti-Asian bias scores on words with varying frequencies in the COHA (1900–1999) corpus (Section 5.3), we compile a novel Asian surname list with the objective of capturing a broader and more representative set of surnames that would be present in a historic corpus such as COHA.

We build on two existing and widely used surname lists for ethnic bias measurement. First, a list of 20 Asian last names curated by Garg et al.

(2018). This was designed to include the most common surnames in the United States for this ethnicity, as measured by 2000 Census data, as well as the surnames that had higher average frequencies in the Google Books and COHA corpora studied by the authors (largely covering the 1800–1999 period). As a result of this curation process, the list is solely focused on higher frequency last names, and, by primarily comprising Chinese surnames, may not be wholly representative of the Asian ethnicity and of the historic appearances of Asian surnames in this corpus. Second, a list of 200 Asian Pacific Islander surnames collected by Matthews et al. (2022) from the 2010 U.S. decennial census, sampled from last names that had a probability of 90% or larger of belonging to this ethnicity.

We expand the set of 211 unique Asian last names from the Garg et al. (2018) and Matthews et al. (2022) lists by collecting surnames in immigration arrival cases from the National Archives at San Francisco, California, from 1910–1940 (National Archives at San Francisco). This data includes over 65,000 cases detailing the country of birth, arrival date, first and last name, gender and date of birth of each person. As ethnicity is not directly reported, we use the country of birth to compile the Asian surname list, and find over 1,300 unique last names belonging to immigrants whose reported birthplace is one the following locations: China, Japan, Korea, Indo-China, India, Hong Kong, Burma, Philippine Islands, Thailand, Malaysia, and Mongolia.[16]

A.4 Alternative Approaches to Word Embedding Estimation Uncertainty

We explored several approaches to measuring the reconstruction error of word embeddings, in addition to the probabilistic model for GloVe that we presented in this work.

A.4.1 Implicit matrix factorization in the skip-gram with negative-sampling (SGNS) model

Per Levy and Goldberg (2014b), the SGNS word embedding model implicitly factorizes a matrix that contains the shifted pointwise mutual information (PMI) of word and context vectors. Let $\mathbf{w}$ be a word vector for word w, $\mathbf{c}$ be a context vector for word c, V_W and V_C be the word and context vocabularies, respectively, k be the number

---

[16]We make this surname list available to researchers in our code repository.
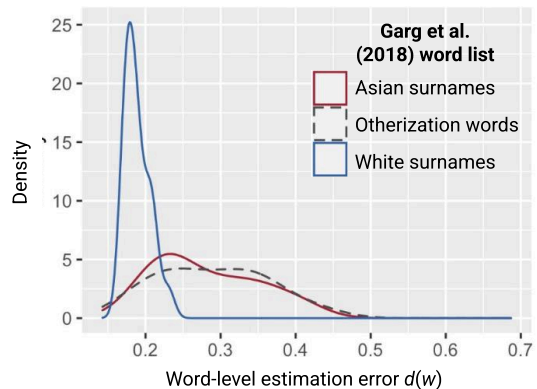

Figure 8: **Distribution of word-level estimation error measures in COHA for different word lists.** Using Levy and Goldberg's findings for the skip-gram with negative-sampling (SGNS) model, we compute a word-level measure of the estimation error in word embeddings.

of negative samples, D be the collection of word and context pairs in a corpus, $\#(i)$ be the number of occurrences of word i in the corpus, and $\#(i, j)$ the number of co-occurrences of the words i and j in the corpus. Then, for sufficiently large dimensionality of the embeddings, the optimal vectors according to the SGNS objective are such that:

$$\mathbf{w} \cdot \mathbf{c} = \log\left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)}\right) - \log k$$
$$= \text{PMI}(w, c) - \log k$$

In this manner, a measure of error in the estimation of the word embedding for word w could be obtained by comparing the dot product $\mathbf{w} \cdot \mathbf{c}$ and $\text{PMI}(w, c)$ across all context words c appearing with word w in the corpus. In particular, we explored a word-level measure of estimation error for word w captured by the median of $d(w, c)$, the context-level percentage of deviation from the optimal value, over all contexts:

$$d(w, c) = \frac{\mathbf{w} \cdot \mathbf{c} - (\text{PMI}(w, c) - \log k)}{\text{PMI}(w, c) - \log k}$$

By comparing the distribution of this word-level measure of estimation error over different word lists, we found that higher frequency word lists, such as the set of White surnames of Garg et al. (2018), had much lower estimation errors compared to relatively lower frequency word lists, such as the set of 20 Asian surnames (see Figure 8).

Though helpful to assess the estimation quality of word embeddings in a corpus, an important limitation of this method is that it is not a probabilistic

model, and so offers no obvious way to perform hypothesis testing in downstream applications accounting for this type of error. For this reason, we sought a different approach that could provide not only point estimates for the embeddings, but accompanying uncertainty measures with a probabilistic foundation.

A.4.2 Bayes by Backprop for the SGNS model

In order to obtain distributions over the word embeddings, we explored the Bayes-by-Backprop algorithm of Blundell et al. (2015), a variational inference approach that learns a distribution over neural network weights. Han et al. (2018) adapt this method to obtain approximate posterior distributions for SGNS word embeddings, incorporating metadata on document covariates in order to learn conditional distributions for different corpus subsets (e.g., temporal periods or genres) that share structural information across these partitions. Using a Gaussian mixture prior for the parameters of the word and context vectors, this method computes the following conditional posterior distribution $\mathbf{w}_{w|x}$ for word vectors and unconditional posterior distribution $\mathbf{w}_c$ for the context vectors:

$$\mathbf{w}_{w|x} \sim N(f(\mu_w, \mu_x), \sigma_{w|c})$$
$$\mathbf{w}_c \sim N(\tilde{\mu}_c, \tilde{\sigma}_c))$$

where x is the subcorpus on which the embedding for word w is estimated, f is an affine transformation that combines corpus-level word vectors μ_w and embeddings for each subcorpora μ_x, and σ_w and $\tilde{\sigma}_c$ are the diagonal covariance of word and context vectors, respectively, parameterized as $\sigma_{w|c} = \log(1 + e^{\rho_w})$ and $\tilde{\sigma}_c = \log(1 + e^{\tilde{\rho}_c})$. The Bayes-by-Backprop algorithm initializes parameters μ_w, μ_x, $\tilde{\mu}_c$, ρ_w and $\tilde{\rho}_c$ for all word and context vectors in the vocabulary, and, given (w, c, x) triplets, performs sequential updates to these parameters by computing the gradient of the variational approximation to the posterior.

Unfortunately, training meaningful embeddings with Bayes-by-Backprop turned out to be challenging. On the COHA corpus, our best performance was a mean accuracy of 0.11 on the Google analogy task (Mikolov et al., 2013a) and a mean Pearson similarity statistic of 0.41 on the MEN similarity task (Bruni et al., 2012), relative to a benchmark of 0.24 and 0.51, respectively, using the pre-trained embeddings of Hamilton et al. (2016) across COHA decades. We explored numerous
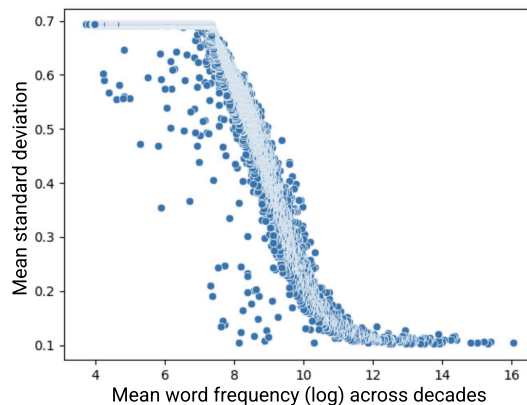

Figure 9: **Word-level relationship between posterior standard deviations and frequency in COHA using the Bayes-by-Backprop approach.** Lower-frequency words displayed poor convergence and their posterior standard deviations were effectively unchanged during training.

refinements to improve the training of Bayesian Neural Networks (Book, 2020), including using different weight initialization schemes for separate parameter groups (e.g., Uniform Kaiming scheme), and both uniformly and dynamically re-weighting the Kullback-Leibler divergence component of the cost function, to no greater success.

In addition to the lower quality of the embedding posterior means, we encountered an important scaling issue in the trained parameters. After training, the posterior means were one to two orders of magnitude smaller than the posterior standard deviations, effectively barring us from drawing meaningful samples from these distributions. An analysis of the relationship between the posterior standard deviations and word-level frequency indicated that there was an inverse relationship between these (similar to what we document in Figure 3 for GloVe-V); however, this relationship only held for higher frequency words. For lower frequency words, the posterior standard deviations were all equivalent and effectively unmodified from their initialization value (see Figure 9). Convergence diagnostics on this model confirmed that these parameters (ρ_w and $\tilde{\rho}_c$) were not training correctly. Further work would be required to design priors and parameter-specific weight initialization schemes that lead to proper training of the parameters for this subset of words. Our code for these training runs is made available at the project repository.