Understanding the Origins of Bias in Word Embeddings (Supplemental Material)

1 Computing the Bias Gradient for GloVe

The bias gradient $\nabla_X B(w(X))$ can be thought of as a $V \times V$ matrix indicating the direction of perturbation of the corpus (co-occurrences) that will result in the maximal change in bias.

$$\begin{aligned}\nabla_X B(w(X)) &= \nabla_w B(w) \nabla_X w(X) \\ &= \sum_{i=1}^{V} \nabla_{w_i} B(w) \nabla_X w_i(X)\end{aligned}$$

Where the first line is obtained through the chain rule, and the second line is a partial expansion of the resulting Jacobian product. Recall $w = \{w_1, w_2, ...w_V\}$, $w_i \in \mathbb{R}^D$ and $X \in \mathbb{R}^{V \times V}$.

When the bias metric is only a function of a small subset of the words in the vocabulary, as in the case of WEAT, this can be further simplified to:

$$\nabla_X B(w(X)) = \sum_{i \in \mathcal{U}} \nabla_{w_i} B(w) \nabla_X w_i(X) \tag{1}$$

Where $\mathcal{U}$ are the indices of the words used by the bias metric; $\mathcal{U} = \mathcal{S} \cup \mathcal{T} \cup \mathcal{A} \cup \mathcal{B}$ for WEAT. For the bias metrics we have explored, the first part of this expression, $\nabla_{w_i} B(w)$, can be efficiently computed through automatic differentiation. The difficulty lies in finding an expression for $\nabla_X w_i(X)$. However, in Section 4.2 of the main text we developed an approximation for the learned embedding under corpus (co-occurrence) perturbations in GloVe using influence functions. We can use this same approximation to create an expression for $w_i(X)$ that is differentiable in X.

Recall, given the learned optimal GloVe parameters w^*, u^*, b^*, c^*, on co-occurrence matrix X, we can approximate the word vectors given a small corpus perturbation as:

$$\tilde{w_i} \approx w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X}_i, w^*) - \nabla_{w_i} L(X_i, w^*) \right] \tag{2}$$

Until now, we have been interested in perturbations stemming from the removal of some part of corpus, e.g. document k, giving us $\tilde{X} = X - X^{(k)}$. However, the above approximation holds for an (almost) arbitrary co-occurrence perturbation, which we shall denote Y. With this change of variable, $\tilde{X} = X - Y$, we can introduce the approximation from Equation (2):

$$\begin{aligned}
\nabla_X w_i(X) &= -\nabla_Y w_i(\tilde{X}(Y))|_{Y=0} \\
&\approx -\nabla_Y \Big[w_i^* - \frac{1}{V} H_{w_i}^{-1} \big[\nabla_{w_i} L(\tilde{X}_i(Y), w^*) - \nabla_{w_i} L(X_i, w^*) \big] \Big] |_{Y=0} \\
&\approx \frac{1}{V} H_{w_i}^{-1} \nabla_Y \nabla_{w_i} L(\tilde{X}_i(Y), w^*)|_{Y=0}
\end{aligned} \tag{3}$$

Where we have made the dependence on Y in Equation (2) explicit. The higher-order jacobian, $\nabla_Y \nabla_{w_i} L(\tilde{X}_i(Y), w^*)|_{Y=0}$, can be thought of as a $D \times V \times V$ tensor. We again note a significant sparsity, since $\tilde{X}_i(Y)$ is only a function of Y_i. Therefore, this tensor is 0 in all but the ith position along one of the V axes. The $D \times V$ "matrix" in that non-zero position can be found by computing:

$$\nabla_{Y_i} \sum_{j=1}^{V} 2V f(X_{ij} - Y_{ij}) \big(w_i^T u_j + b_i + c_j - \log(X_{ij} - Y_{ij}) \big) u_j$$

evaluated at $Y_{ij} = 0$. Alternatively the Jacobian can simply by obtained using automatic differentiation.

Substituting this result into Equation (1), we get:

$$\begin{aligned}
\nabla_X B(w(X)) &= \sum_{i \in \mathcal{U}} \nabla_{w_i} B(w) \nabla_X w_i(X) \\
&\approx \frac{1}{V} \sum_{i \in \mathcal{U}} \nabla_{w_i} B(w) H_{w_i}^{-1} \nabla_Y \nabla_{w_i} L(\tilde{X}_i(Y), w^*)|_{Y=0}
\end{aligned}$$

Which gives us the full approximation of the Bias Gradient in GloVe.

Note that since $\nabla_{w_i} L(\tilde{X}_i(Y), w^*)$ is not differentiable in Y at $Y = 0$ where $X_{ij} = 0$, the bias gradient is only defined at non-zero co-occurrences. This prevents us from using the bias gradient to study corpus additions which create previously unseen word co-occurrences. However, this does not affect our ability to study arbitrary removals from the corpus, since removals cannot affect a zero-valued co-occurrence. Of course, nothing limits us from using the bias gradient to also consider additions to the corpus that not change the set of zero co-occurrences.

2 Experimental Setup

Table 1 presents a summary of the corpora and embedding hyperparameters used throughout our experimentation. We list the complete set of the words used in each of the two WEATs below.

Table 1: Experimental Setups

	Wiki	**NYT**
Corpus		
Min. doc. length	200	100
Max. doc. length	10,000	30,000
Num. documents	29,344	1,412,846
Num. tokens	17,033,637	975,624,317
Vocabulary		
Token min. count	15	15
Vocabulary size	44,806	213,687
GloVe		
Context window	symmetric	symmetric
Window size	8	8
α	0.75	0.75
x_{max}	100	100
Vector Dimension	75	200
Training epochs	300	150
Performance		
TOP-1 Analogy	35%	54%

WEAT 1

$\mathcal{S}$	**science**	science, technology, physics, chemistry, einstein, nasa, experiment, astronomy
$\mathcal{T}$	**arts**	poetry, art, shakespeare, dance, literature, novel, symphony, drama
$\mathcal{A}$	**male**	male, man, boy, brother, he, him, his, son
$\mathcal{B}$	**female**	female, woman, girl, sister, she, her, hers, daughter

WEAT 2

$\mathcal{S}$	**instruments**	bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin
$\mathcal{T}$	**weapons**	arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip
$\mathcal{A}$	**pleasant**	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation
$\mathcal{B}$	**unpleasant**	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

3 Detailed Experimental Methodology

Here we detail the experimental methodology used to test our method's accuracy.

***I* - Train a baseline.** We start by training 10 word embeddings using the parameters in Table 1 above, but using different random seeds. These embeddings create a baseline for the unperturbed bias $B(w^*)$.

***II* - Approximate the differential bias of each document.** For each WEAT test, we approximate the differential bias of every document in the corpus. We do so with a combination of Equations (8) and (5) of the main text. This step is summarize by Algorithm 1 in the main text. Note that we make the differential bias approximation for each document several times, using the learned parameters w^*, u^*, b^* and c^* from the 10 different baseline embeddings in our different approximations. We then average these approximations for each document, and construct a histogram.

***III* - Construct perturbation sets.** We perturb the corpus by removing sets of documents. We construct three types of perturbation sets: *increase*, *random*, and *decrease*. The targeted (increase, decrease) perturbation sets are constructed from the documents whose removals were predicted to cause the greatest differential bias (in absolute value), i.e., the documents located in the tails of the histograms. For the Wiki setup we consider the 10, 30, 100, 300, and 1000 most influential documents for each bias, while for the NYT setup we consider the 100, 300, 1000, 3000, and 10,000 most influential. This results in 10 perturbations sets per corpus per bias, for a total of 40.

The random sets are, as their name suggests, drawn uniformly at random from the entire set of documents used in the training corpus. For the Wiki setup we consider 6 sets of 10, 30, 100, 300, and 1000 documents (30 total). Because training times are much longer, we limit this to 6 sets of 10,000 documents for the NYT setup. Therefore we consider a total of 36 random sets.

***IV* - Approximate the differential bias of each perturbation set.** We then approximate the differential bias of each perturbation set. Note that $\nabla_w L(X_i, w)$ is not linear in X_i. Therefore determining the differential bias of a perturbation set does not amount to simply summing the differential bias of each document in the set (although in practice we find it to be close). Here we also make 10 approximations, one with each of the different baseline embeddings.

***V* - Construct ground truth and assess.** Finally, for each perturbation set, we remove the target documents from the corpus, and train 5 new embeddings on this perturbed corpus. We use the same hyperparameters, again varying only the random seed. The bias measured in these embeddings serve as the ground truth for assessment.

4 Additional experimental results

Here we include additional experimental results.

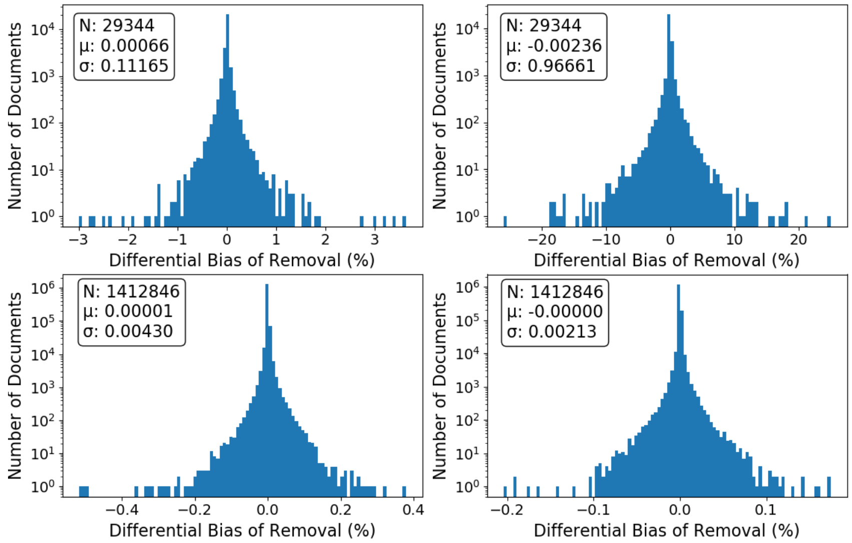

Figure 1: Histogram of the approximated differential bias of removal for every document in our Wiki setup (top) and NYT setup (bottom), considering WEAT1 (left) and WEAT2 (right), measured in percent change from the corresponding mean baseline bias.

Table 2: Correlation of Approximated and Validated Mean Biases

	WEAT1	WEAT2
Wiki	r^2: 0.986	r^2: 0.993
NYT	r^2: 0.995	r^2: 0.997

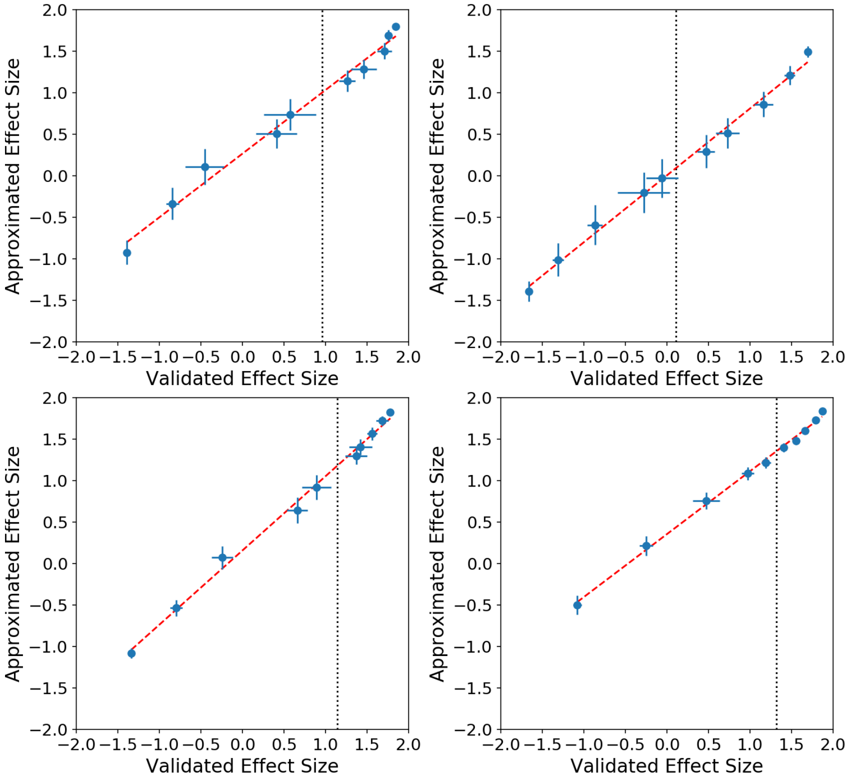

Figure 2: Approximated vs. ground truth WEAT bias effect size due to the removal of each (non-random) perturbation set in Wiki setup (top) and NYT setup (bottom), considering WEAT1 (left) and WEAT2 (right); points plot the means; error bars depict one standard deviation; dashed line shows least squares; the baseline means are shown with vertical dotted lines; correlations in Table 2.

Table 3: A comparison of the effect of removing the most impactful documents as identified by a PPMI baseline technique versus when identified by our method (Wiki setup, mean of WEAT1 in 10 retrained GloVe embeddings).

Document Set		ΔB when Identified by	
objective	num. docs.	baseline	our method
correct	300	-67%	-187%
correct	100	-50%	-147%
correct	30	-23%	-57%
correct	10	-4%	-40%
aggravate	10	-0.5%	32%
aggravate	30	20%	53%
aggravate	100	15%	79%
aggravate	300	47%	84%

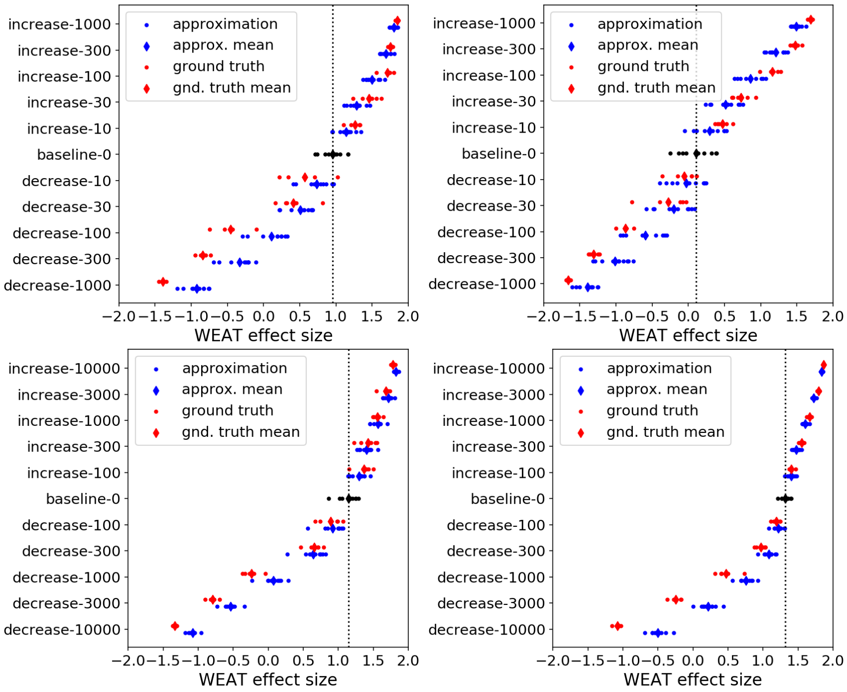

Figure 3: Approximated and ground truth differential bias of removal for every (non-random) perturbation set in Wiki setup (top) and NYT setup (bottom), considering WEAT1 (left) and WEAT2 (right); the baseline means are shown with vertical dotted lines

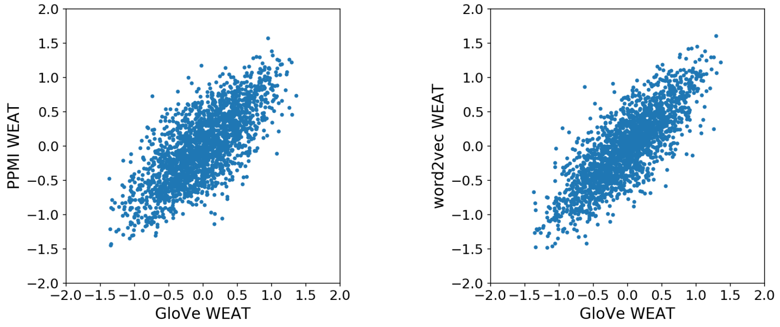

Figure 4: The correlation of the WEAT as measured in our NYT GloVe embeddings versus the corpus' PPMI representation in 2000 randomly generated word sets, $r^2 = 0.725$ (left); versus when measured in word2vec embeddings with comparable hyper-parameters, $r^2 = 0.803$ (right).

5 Influential Documents - NYT WEAT 1

The below documents were identified to be the 50 most WEAT1 bias influencing documents in our NYT setup. We list the article titles. Publication dates range from January 1, 1987 to June 19, 2007. Most can be found through *https://www.nytimes.com/search*. A subscription may be required for access.

$\Delta_{\text{doc}}B$	**Bias Decreasing**
-0.52	Hormone Therapy Study Finds Risk for Some
-0.50	For Women in Astronomy, a Glass Ceiling in the Sky
-0.49	Sorting Through the Confusion Over Estrogen
-0.36	Young Astronomers Scan Night Sky and Help Wanted Ads
-0.33	Campus Where Stars Are a Major
-0.33	A New Look At Estrogen And Stroke
-0.31	Scenes From a Space Thriller
-0.30	The Cosmos Gets Another Set of Eyes
-0.29	The Stars Can't Help It
-0.28	Making Science Fact, Now Chronicling Science Fiction
-0.27	Estrogen Heart Study Proves Discouraging
-0.26	EINSTEIN LETTERS TELL OF ANGUISHED LOVE AFFAIR
-0.25	Divorcing Astronomy
-0.24	Astronomers Open New Search for Alien Life
-0.23	AT WORK WITH: Susie Cox; Even Stars Need a Map To the Galaxy
-0.22	CAMPUS LIFE: Minnesota; Astronomer Spots Clue To Future of Universe
-0.21	Clothes That Are Colorful and TV's That Are Thin Make Many Lists
-0.20	We Are the Fourth World
-0.20	Hitched to a Star, With a Go-To Gadget
-0.19	Material World
-0.19	Shuttle's Stargazing Disappoints Astronomers
-0.19	2 Equity Firms Set to Acquire Neiman Marcus
-0.18	Volunteer's Chain Letter Embarrasses a Hospital
-0.18	What Doctors Don't Know (Almost Everything)
-0.18	Astronomers Edging Closer To Gaining Black Hole Image

$\Delta_{\text{doc}}B$	**Bias Increasing**
0.38	Kaj Aage Strand, 93, Astronomer At the U.S. Naval Observatory
0.32	Gunman in Iowa Wrote of Plans In Five Letters
0.29	ENGINEER WARNED ABOUT DIRE IMPACT OF LIFTOFF DAMAGE
0.29	Fred Gillett, 64; Studied Infrared Astronomy
0.27	Robert Harrington, 50, Astronomer in Capital
0.27	For Voyager 2's 'Family' of 17 Years, It's the Last of the First Encounters
0.26	Despite the Light, Astronomers Survive
0.25	LONG ISLAND GUIDE
0.25	THE GUIDE
0.24	Telescope Will Offer X-Ray View Of Cosmos
0.23	Astronomers Debate Conflicting Answers for the Age of the Universe
0.23	The Wild Country of Anza Borrego
0.21	What Time Is It in the Transept?
0.21	Jan H. Oort, Dutch Astronomer In Forefront of Field, Dies at 92
0.21	Logging On to the Stars
0.20	The Sky, Up Close and Digital
0.20	Q&A
0.19	Getting Attention With Texas Excess
0.19	Emily's College
0.19	60 New Members Elected to Academy of Sciences
0.18	Theoretical Physics, in Video: A Thrill Ride to 'the Other Side of Infinity'
0.18	Charles A. Federer Jr., Stargazer-Editor, 90
0.18	Some Web sites are taking their brands from the Internet into some very offline spheres.
0.18	A Wealth of Cultural Nuggets Waiting to Be Mined
0.18	Can a Robot Save Hubble? More Scientists Think So

6 Influence of Mulitple Perturbations

Here we show how we can extend the influence function equations presented by Koh & Liang (2017) to address the case of multiple training point perturbations. We do not intend this to be a rigorous mathematical proof, but rather to provide insight into the logical steps we followed.

First we summarize the derivation in the case of a single train point perturbation. Let $R(z, \theta)$ be a convex scalar loss function for a learning task, with optimal model parameters θ^* of the form in Equation 4 below, where $\{z_1, ..., z_n\}$ are the training data points and $L(z_i, \theta)$ is the point-wise loss.

$$R(z,\theta) = \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) \qquad\qquad \theta^* = \underset{\theta}{\operatorname{argmin}}\, R(z,\theta) \tag{4}$$

We would like to determine how the optimal parameters θ^* would change if we perturbed the k^{th} point in the training set; i.e., $z_k \to \tilde{z}_k$. The optimal parameters under perturbation can be written as:

$$\tilde{\theta}(\varepsilon) = \underset{\theta}{\operatorname{argmin}} \left\{ R(z,\theta) + \varepsilon L(\tilde{z}_k,\theta) - \varepsilon L(z_k,\theta) \right\} \tag{5}$$

where we seek $\tilde{\theta}|_{\varepsilon=\frac{1}{n}}$, noting that $\tilde{\theta}|_{\varepsilon=0} = \theta^*$. Since $\tilde{\theta}$ minimizes Equation 5, we must have

$$0 = \nabla_\theta R(z,\tilde{\theta}) + \varepsilon \nabla_\theta L(\tilde{z}_k,\tilde{\theta}) - \varepsilon \nabla_\theta L(z_k,\tilde{\theta})$$

for which we can compute the first order Taylor series expansion (with respect to θ) around θ^*. This gives:

$$\begin{aligned} 0 \approx& \nabla_\theta R(z,\theta^*) + \varepsilon \nabla_\theta L(\tilde{z}_k,\theta^*) - \varepsilon \nabla_\theta L(z_k,\theta^*) \\ &+ \left[\nabla_\theta^2 R(z,\theta^*) + \varepsilon \nabla_\theta^2 L(\tilde{z}_k,\theta^*) - \varepsilon \nabla_\theta^2 L(z_k,\theta^*)\right](\tilde{\theta} - \theta^*) \end{aligned}$$

Noting $\nabla_\theta R(z,\theta^*) = 0$, then keeping only $\mathrm{O}\big(\varepsilon\big)$ terms, solving for $\tilde{\theta}$, and evaluating at $\varepsilon = \frac{1}{n}$ we obtain:

$$\tilde{\theta} - \theta^* \approx \left(\frac{-1}{n}\right) H_{\theta^*}^{-1} \left[\nabla_\theta L(\tilde{z}_k,\theta^*) - \nabla_\theta L(z_k,\theta^*)\right] \tag{6}$$

where $H_{\theta^*} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i,\theta^*)$ is the Hessian of the total loss.

Now, we address the more general case where several training points are perturbed. This corresponds to replacing the expression $\varepsilon L(\tilde{z}_k,\theta) - \varepsilon L(z_k,\theta)$ in Equation (5) with $\sum_{k\in\delta} \big(\varepsilon L(\tilde{z}_k,\theta) - \varepsilon L(z_k,\theta)\big)$, where δ is the set of indices of perturbed points. Because of the linearity of the gradient operator, we can readily carry this substitution through the subsequent equations, resulting in:

$$\tilde{\theta} - \theta^* \approx \left(\frac{-1}{n}\right) H_{\theta^*}^{-1} \sum_{k\in\delta} \left[\nabla_\theta L(\tilde{z}_k,\theta^*) - \nabla_\theta L(z_k,\theta^*)\right] \tag{7}$$

where we assume $|\delta| \ll n$.