Posterior Sampling of Word Embeddings

Väinö Yrjänäinen*, Isac Boström†, Johan Jonasson†, Måns Magnusson*

*Department of Statistics, Uppsala University, Uppsala, Sweden.
†Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden.

Abstract

Quantifying uncertainty in word embeddings is crucial for reliable inference from textual data, yet existing methods like bootstrap and mean-field variational inference (MFVI) are computationally intensive or make limiting assumptions. We explore alternative approaches, focusing particularly on Gibbs sampling via Polya-Gamma augmentation as our key contribution, alongside Laplace approximation and Hamiltonian Monte Carlo (HMC). Additionally, we address the challenge of non-identifiability in word embeddings. In simulation studies with known ground truth, our methods demonstrate superior performance in capturing true uncertainties compared to existing approaches.

Skip-gram negative samples

We develop uncertainty estimation for probabilistic word embeddings (Rudolph et al., 2016; Bamler and Mandt, 2017; Rudolph and Blei, 2017), which are probabilistic formulations of the skip-gram with negative samples (SGNS) model first presented by Mikolov et al. (2013).

Let $\mathcal{D} = (w_i)_{i=1}^N$ denote the dataset. For each word $w_j \in W = \{w_j\}_{j=1}^V$ we associate two embedding vectors: A *target vector* $\rho_j \in \mathbb{R}^K$ and a *context vector* $\alpha_j \in \mathbb{R}^K$. Where K denotes the dimensionality of the embedding. We organize these into matrices, $\rho \in \mathbb{R}^{V\times K}$ and $\alpha \in \mathbb{R}^{V\times K}$. The complete set of parameters $\theta \in \mathbb{R}^{2V\times K}$ is then defined as $\theta = [\rho, \alpha]^\top$.

The SGNS likelihood nicely factors into terms, where each term only has a handful of parameters

$$\log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \Bigg(\underbrace{\sum_{v\in C_i^+} \log \sigma(\rho_{w_i}^T \alpha_v)}_{\text{positive samples}} + \underbrace{\sum_{v\in C_i^-} \log(1-\sigma(\rho_{w_i}^T \alpha_v))}_{\text{negative samples}} \Bigg) \tag{1}$$

where C_i^+ is the context window, or the set of positive samples, for word w_i. The negative samples C_i^- are drawn randomly from the empirical distribution.

Embeddings and rotations

The SGNS likelihood, as presented in Equation 1, fully consists of dot products between the target vectors ρ_v and the context vectors α_v. Given any invertible linear transformation $\boldsymbol{A} \in \mathrm{GL}(K)$, we can apply the following paired transformations to the embeddings: $\rho'_w = \boldsymbol{A}\rho_w$ and $\alpha'_v = \boldsymbol{A}^{-\top}\alpha_v$. Under these transformations the dot product remains invariant, $(\rho'_w)^\top \alpha'_v = (\rho_w)^\top \alpha_v$. And thus the likelihood remains unchanged.

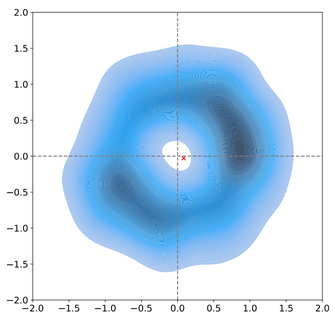


Figure 1: $\rho_{1,1}$ and $\rho_{1,2}$ display a donut-like symmetry. $D = 2$, $N = 20000$ (as in all plots where not explicitly stated).

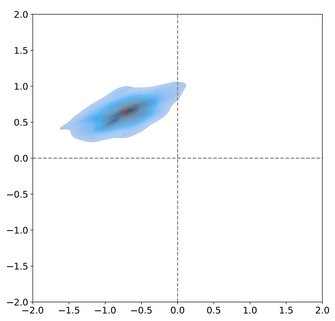


Figure 2: Fixing $K \times K$ elements eliminates the rotational symmetries for the posterior distribution of $\rho_{1,1}$ and $\rho_{1,2}$.

To eliminate the symmetries, $K \times K$ parameters need to be fixed. We propose having a reference MAP estimate, from which K context vectors are selected, and their values are fixed in the subsequent estimation algorithm. The resulting posterior is demonstrated in Figure 2.

Gibbs sampler

The distribution of each ρ_v for any $v \in W$ is the same as the posterior for a logistic regression. Utilizing that, we build a Gibbs sampler that alternates sampling ρ and α. Moreover, we use the Polya-Gamma method by Polson et al. (2013) to sample from the logistic posterior.

1: **function** Embedding Gibbs sampler(x, W, N, D, warmup)
2: ▷ Initialize ρ and α
3: **for** $w \in W$ **do**
4: $\rho_w \sim \mathcal{N}_D\left(0, \frac{1}{\sqrt{D}}\boldsymbol{I}\right)$
5: $\alpha_w \sim \mathcal{N}_D\left(0, \frac{1}{\sqrt{D}}\boldsymbol{I}\right)$
6:
7: **for** $i \in N$ **do**
8: **for** $w \in W$ **do**
9: $\rho_w \leftarrow$ **Polya-Gamma-Logistic** (α, x)
10: **for** $w \in W$ **do**
11: $\alpha_w \leftarrow$ **Polya-Gamma-Logistic** (ρ, x)
12: **if** $i \geq$ warmup **then**
13: ▷ Yield the sample on each iteration
14: **return** ρ, α

Since each ρ_v and α_v is conditionally independent given α and ρ, respectively, the algorithm can be highly parallellized. Specifically, the for loops on rows 8 and 10 can be completely parallellized, enabling a theoretical $|W|$ fold speedup.

Experiments

To be able to compare embeddings with different rotations, we define the probability divergence between embeddings θ and θ' as the root mean squared error between the co-occurence probabilities

$$d_{co}(\theta, \theta') = \sqrt{\frac{1}{V^2} \sum_{v,w\in W} (P(w \wedge v \mid \theta) - P(w \wedge v \mid \theta'))^2} \tag{2}$$

As a first step, we measure the convergence of the methods to the true values given the simulated data. For each $w \in W$, the word and context vectors are simulated from

$$\begin{aligned} \rho_w &\sim \mathcal{N}(0, \varepsilon^2 I/K), \\ \alpha_w &\sim \mathcal{N}(0, \varepsilon^2 I/K), \end{aligned} \tag{3}$$

where d is the dimensionality of the embedding, and the hyperparameter $\varepsilon = 1$. We generate random word pairs (w, v) by sampling uniformly from the set W. A Bernoulli random variable X is then sampled

$$\begin{aligned} w &\sim \mathrm{Uniform}(W) \\ v &\sim \mathrm{Uniform}(W) \\ X &\sim \mathrm{Bernoulli}(\sigma(\alpha_v^T \rho_w)) \end{aligned} \tag{4}$$

N times for a simulated dataset with N observations.

Figure 3 plots convergence of the different estimation methods in terms of d_{co} to the true parameters. No large differences across the estimation methods are observed. Specifically, the Gibbs sampler and HMC yield very similar results.

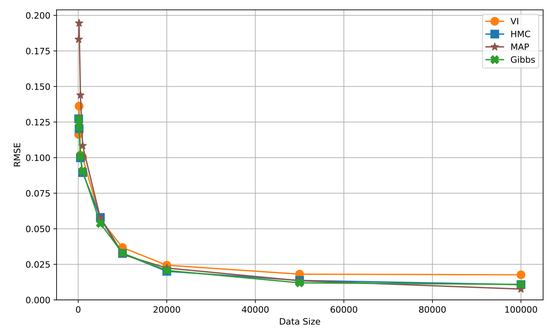


Figure 3: Convergence with the different estimation methods. d_{co} (RMSE) as a metric.

We also study the co-occurence probabilities of specific word pairs and in particular the 90% credible interval. This is illustrated for one word-pair in Figure 4, and the percentage of coverage of all word-pairs are presented in Table 1.

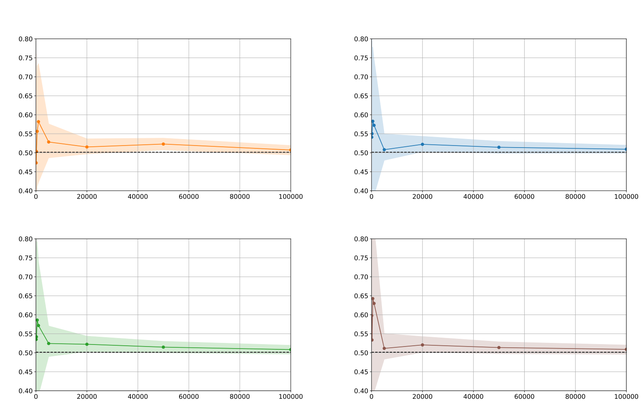
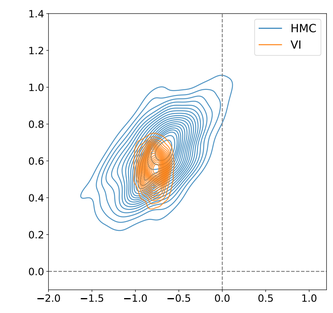


Figure 4: Example of co-occurence probability with increasing data size, using different estimation methods. The methods are presented in the following order: MFVI (orange), HMC (blue), Gibbs (green), Laplace (brown). The bold lines represent the mean of the co-occurence probability and the surrounding faded region is the 90% credible interval. In addition, the black dashed lines represents the true co-occurence probability.

Table 1: Coverage (%) of the true co-occurence probability of the 90%-credible interval in the simulated experiment for different data sizes. Averaged over 10 separate simulation experiments.

Gibbs	91.1	83.3	88.9	90.0	91.1
Laplace	92.0	90.2	87.8	85.7	86.2
HMC	91.9	88.8	88.1	85.8	86.8
MFVI	86.2	84.0	80.2	78.8	61.1
	1000	5000	20000	50000	100000



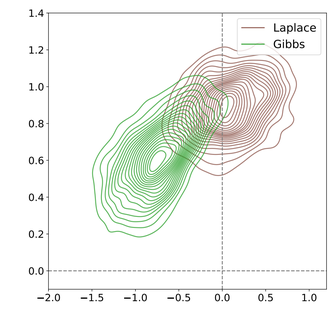


Figure 5: Posterior for $\rho_{1,1}$ and $\rho_{1,2}$ for the different methods. Rotations are eliminated by fixing $D \times D$ parameters to a MAP estimate.

In Figure 5, the posterior for specific parameter values is shown. The parameter values are now comparable, as the rotations are eliminated. Mean-field variational inference seems to underestimate the uncertainty of the parameters compared to the other methods, which also seen in Table 1. This is in line with previous literature on variational inference (Wang and Blei, 2018).

References

Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Maja Rudolph and David Blei. Dynamic Bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*, 2017.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. Exponential family embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

Yixin Wang and David M Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 2018.