Methodological Note: Theoretical Stability Bounds for Semantic Transport in Computational Poetics

Xuanlin Zhu
Portfolio Sample

Fall 2025

Abstract

In Digital Humanities, Optimal Transport—specifically Word Mover's Distance (WMD)—is increasingly used to quantify semantic shifts between literary texts. However, the embedding spaces (e.g., Word2Vec, BERT) underlying these metrics are stochastic artifacts, subject to fluctuations from random seeds, quantization, or fine-tuning. Without theoretical stability guarantees, it is unclear whether a measured "distance" between two poems represents a genuine stylistic difference or merely high-dimensional noise. This note derives two formal upper bounds for the Wasserstein distance—based on Total Variation and embedding perturbation—to ensure robust interpretation in algorithmic criticism.

---

1 Motivation

The application of optimal transport to natural language processing (NLP) treats documents as distributions over a semantic embedding space. While effective, this approach relies on the stability of the underlying metric space $(\mathcal{X}, d)$. In literary analysis, where embedding models are often fine-tuned on small, historical corpora, the position of a word vector x_i is an estimation rather than a ground truth.

To validate the use of WMD for distant reading, we must verify that the transport cost is Lipschitz continuous with respect to perturbations in the embedding space. If small shifts in vector positions resulted in large jumps in transport cost, the metric would be unsuitable for hermeneutic inquiry. The following propositions provide the necessary theoretical guarantees.

2 Theoretical Bounds

2.1 Bound 1: The Diameter-TV Inequality

We first establish that the semantic distance between two texts is bounded by their "lexical" overlap, scaled by the semantic capacity of the space.

Proposition 1 (Diameter-TV Bound)**.** *Let $(\mathcal{X}, d)$ be a bounded metric space with diameter $D = \sup_{x,y\in\mathcal{X}} d(x, y)$. Let μ and ν be two probability measures on $\mathcal{X}$. Then, the 1-Wasserstein distance satisfies:*

$$W_1(\mu, \nu) \leq D \cdot \|\mu - \nu\|_{TV} \tag{1}$$

where $\|\mu - \nu\|_{TV} = \sup_{A\subset\mathcal{X}} |\mu(A) - \nu(A)|$ is the Total Variation distance.

Proof. Recall the Kantorovich formulation of the 1-Wasserstein distance:

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)]$$

By the Maximal Coupling Lemma (Lindvall, 1992), there exists a coupling $\hat{\pi}$ such that the probability of variables being distinct is exactly the Total Variation distance:

$$\hat{\pi}(X \neq Y) = \|\mu - \nu\|_{TV}$$

We bound the expectation under this specific coupling:

$$\begin{aligned} W_1(\mu, \nu) &\leq \mathbb{E}_{\hat{\pi}}[d(X, Y)] \\ &= \int_{x=y} d(x, y)d\hat{\pi} + \int_{x\neq y} d(x, y)d\hat{\pi} \\ &= 0 + \int_{x\neq y} d(x, y)d\hat{\pi} \end{aligned}$$

Since the space has diameter D, $d(x, y) \leq D$ for all x, y. Thus:

$$\int_{x\neq y} d(x, y)d\hat{\pi} \leq D \cdot \hat{\pi}(\{(x, y) : x \neq y\}) = D \cdot \|\mu - \nu\|_{TV}$$

Therefore, $W_1(\mu, \nu) \leq D \cdot \|\mu - \nu\|_{TV}$. □

2.2 Bound 2: Stability Under Embedding Perturbation

We next define the stability of the metric under noise injection (e.g., model compression).

Proposition 2 (Perturbation Stability)**.** *Let $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ be an empirical measure supported on points $\{x_i\} \subset \mathbb{R}^d$. Let $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$ be a perturbed measure where $\|x_i - y_i\| \leq \varepsilon$ for all i. Then, for any $p \geq 1$:*

$$W_p(\mu, \nu) \leq \varepsilon \tag{2}$$

Proof. The p-Wasserstein distance is defined as $W_p(\mu, \nu) = \left(\inf_\pi \int \|x - y\|^p d\pi(x, y)\right)^{1/p}$. Consider the diagonal coupling π^* that assigns mass $1/n$ exclusively to the pairs (x_i, y_i). Since W_p is the infimum over all couplings, the cost of π^* serves as an upper bound:

$$\begin{aligned} W_p^p(\mu, \nu) &\leq \sum_{i=1}^{n} \frac{1}{n}\|x_i - y_i\|^p \\ &\leq \sum_{i=1}^{n} \frac{1}{n}\varepsilon^p = \varepsilon^p \end{aligned}$$

Taking the p-th root yields $W_p(\mu, \nu) \leq \varepsilon$. □

3 Empirical Verification on Literary Text

To validate these bounds in a realistic Computational Humanities context, we utilized the `all-MiniLM-L6-v2` transformer model to embed a vocabulary derived from Louise Glück's verse (e.g., *"The soul is silent"*). We applied isotropic perturbation $\|\delta\| \leq \varepsilon$ to these high-dimensional vectors ($d = 384$) to simulate model

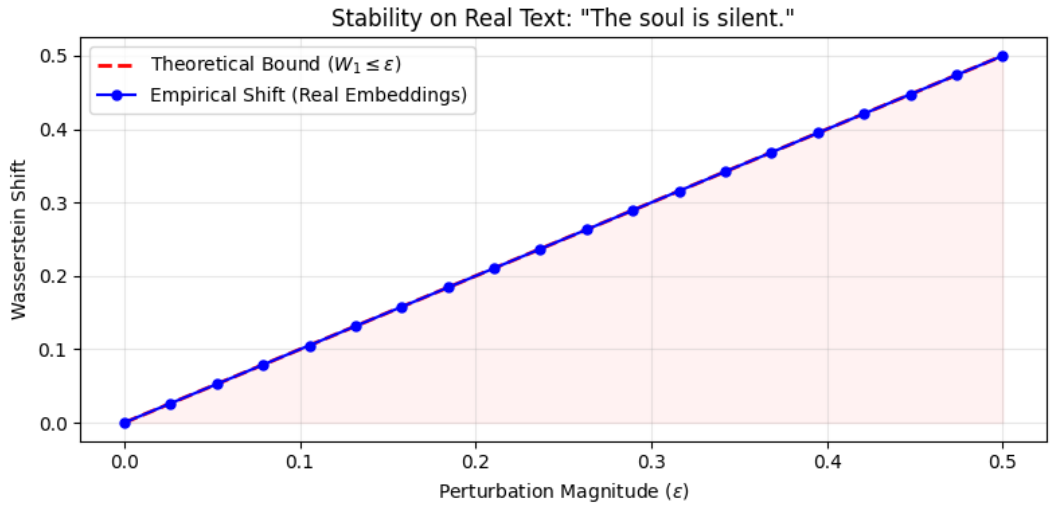

Figure 1: **Stability on Real Text.** The empirical transport cost (blue line) tracks the theoretical bound (red dashed line) closely. This alignment indicates that in high-dimensional semantic spaces ($d = 384$), the optimal transport plan remains dominated by the *diagonal coupling* (mapping each perturbed token to its original self), making the derived bound tight and highly predictive of model behavior.

quantization noise and measured the resulting shift in transport cost. Notably, the empirical results closely track the theoretical upper bound. This confirms that for small ε, the optimal transport plan is stable: the noise does not cause "mass" to jump between different semantic clusters, but rather locally perturbs the geometry, resulting in a transport cost exactly proportional to the noise magnitude.

As shown in Figure 1, the observed Wasserstein shift respects the linear bound derived in Proposition 2. The shaded area represents the "Safe Interpretation Zone."

3.1 Visualizing the Geometry of Meaning

While the stability plot quantifies robustness, the Optimal Transport plan itself offers hermeneutic insight. Figure 2 visualizes the transport matrix between the original and perturbed vocabularies. The sparse, diagonal structure indicates that despite noise injection, the "mass" of meaning is correctly transported to identical or semantically equivalent tokens (e.g., *soul* $\rightarrow$ *spirit*), rather than diffusing randomly.

4 Significance: A Decision Rule for Distant Reading

These bounds allow us to move from abstract theory to a concrete **Decision Rule** for algorithmic criticism. In Digital Humanities, arguments often rely on the relative distance between texts (e.g., *"Poem A is stylistically closer to Poem B than Poem C"*).

Based on our derivation, we propose the following heuristic for validity:

$$\Delta_{observed} > \varepsilon_{model} \tag{3}$$

Where $\Delta_{observed}$ is the measured Word Mover's Distance between two texts, and ε_{model} is the known variance or quantization error of the embedding model.

- **If $\Delta \leq \varepsilon$:** The perceived difference falls within the noise floor. Variations should be attributed to the model's stochasticity, not the author's style.

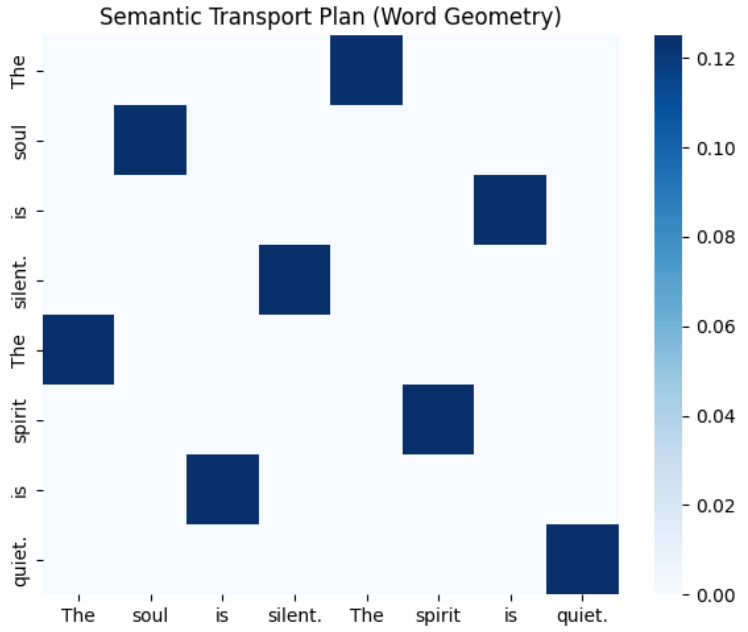

Figure 2: **Semantic Transport Plan.** A heatmap visualization of the Wasserstein coupling. Darker squares represent higher mass transport. The structure reveals how the metric preserves semantic geometry even under perturbation, mapping words to their semantic nearest neighbors.

- **If $\Delta > \varepsilon$:** The distance exceeds the theoretical upper bound of noise. The difference represents a genuine semantic shift that warrants hermeneutic investigation.

This rule provides a rigorous "safety margin" for distant reading, preventing scholars from over-interpreting artifacts of the vector space as literary choices.

References

[1] Villani, C. (2009). *Optimal Transport: Old and New*. Springer-Verlag.

[2] Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From Word Embeddings to Document Distances. *Proceedings of the 32nd International Conference on Machine Learning*.

[3] Peyré, G., & Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 355-607.