Reviewer #1: Our key contribution is to explain mathematically how word embeddings of Glove/W2V capture semantic properties of words. Whilst some aspects relate to previous empirical observations or hypotheses e.g. a connection between relatedness and PMI (as acknowledged), to our knowledge no previous work *theoretically*: (i) explains the semantic properties of W2V/Glove embeddings as following from their low-rank projections of PMI vectors (S4); (ii) explores geometric properties of the space of PMI vectors and thus of word embeddings (S4.1); (iii) proves that relatedness, similarity, paraphrase and analogy (as defined) each correspond to mathematical relationships of PMI vectors with defined error terms that, e.g., in part explain the variability observed in analogy relationships [e.g. 22, Rogers et al. (2017)] (S4.2-4.4); (iv) concludes that those semantic relationships are best preserved by linear projection, under which embedding interactions can be probabilistically interpreted (S5) and several well-known observations/heuristics explained (S6); and (vi) from their formulations, derive a mathematical connection between relatedness, similarity, paraphrase and analogy. Regarding the referenced "observations" and "specifics":

- In S5.1, we prove $\mathbf{W} \neq \mathbf{C}$ since the opposite has been assumed [3,13]. Our result explains why tying $\mathbf{W} = \mathbf{C}$ gives good but sub-optimal results (since most eigenvalues are positive) and enables that sub-optimality to be quantified.
- In S5.2, we explicitly define the inherent error of low-rank approximation and the additional error due to taking the dot-product of embeddings from $\mathbf{W}$, as is often the case. We also quantify the effect of average embeddings.
- In S4, *similarity* corresponds to the “attributional similarity” of [37] and *paraphrasing* is as defined in [11, 2]. As such, “w_a is similar to w_b” is equivalent to “w_a paraphrases $\mathcal{W} = \{w_b\}$”. We will make this more clear.
- In S4.2, we show that *subtraction* of PMI vectors equates to un-weighted KL divergence, thus where a PMI vector difference is small, the KL divergence is small and words are similar. This mathematically proves that (and how) PMI-based word embeddings instantiate the hypothesis that "similar row vectors in the word-context matrix indicate similar word meanings" for a general word-context matrix and unspecified vector-similarity measure [37].
- S4.3 shows how *addition* of PMI vectors (for words in a set $\mathcal{W}$) corresponds to identifying a paraphrase of words in $\mathcal{W}$, subject to their mutual independence, from which we can geometrically interpret the difference between the sum of PMI vectors and the PMI vector of the paraphrase word.
- S4 (beginning) justifies the view of word embeddings as projections of PMI vectors by extending [19]; S5 (beginning) motivates the use of linear projection. We believe both of these are original perspectives for W2V/Glove.
- The *LSQ* model implements loss function (11), details are in S6 (end) and Appendix F. We will include that all implementations use PyTorch with the Adam optimiser and review wording to ensure it is comprehensive and clear.

Many works investigate W2V/Glove and their embeddings [18, 22, 4, 13, 3, 8, 17, 11, 2, 9], indicating that they do “remain poorly understood” [28]. We provide, *inter alia*, a first explicit mathematical understanding of what W2V/Glove embedding parameters represent and the semantic relationships they capture. We believe our work is relevant as these algorithms have been adopted by other domains [30, 31], their embeddings are ubiquitous and the presence of linear relationships between word embeddings has been recently questioned (Rogers et al. (2017), Schluter (2018)). We provide interpretability of W2V and Glove, enabling principled improvement of word embeddings (as shown with LSQ loss) and comparison metrics; and interpretability in down-stream tasks. We will make our contributions more clear.

Reviewer #2: *Global relatedness* (GR) refers to when two words (or word sets) induce similar distributions over all other words, which underpins the definitions of similarity, paraphrase and analogy. GR manifests geometrically as a small difference between one PMI vector (or sum of PMI vectors) and another, meaning that the associated KL divergence is small and the relevant semantic relationship (similarity, paraphrase or analogy) exists. Since the interactions of PMI vectors associated with those semantic relationships are linear, when PMI vectors are projected linearly to a lower dimension, those relationships are necessarily maintained and GR remains identifiable between the low-dimensional representations, i.e. word embeddings. In summary, as a kind of equivalence measure within semantic relationships, global relatedness corresponds to small vector difference in PMI space and thus also in the space of word embeddings given a sufficiently homomorphic (e.g. linear) projection. We will make this more clear in the paper.

Reviewer #3: Section 5.2 (end) considers the interpretation of cosine similarity. While cosine similarity is not found to have an obvious probabilistic interpretation, we conjecture (based on the other mathematical relationships derived) that it serves as a blended measure of similarity and relatedness. The dot product numerator approximates relatedness, but, as you say, can be high even if vectors are far apart in Euclidean distance (e.g. words that are related but not similar). By normalising, cosine similarity requires the angle between embeddings to be small, which is closer to a similarity test, thus cosine similarity appears to evaluate somewhere between relatedness and similarity, explaining why it has been used to measure both [33, 4]. To clarify, the difference between two PMI vectors is indeed given by ρ of Eq. 8 (note: ϵ of Eq. 9 applies only when PMI vectors are added to find paraphrases). In PMI space, an appropriately weighted sum over ρ components gives a KL divergence and thus a probabilistic measure of word similarity. We suggest (S5.2) that, by dropping low probability dimensions, the low-dimension projection to word embeddings approximates such a probabilistic weighting over dimensions. In future work we plan to investigate these interactions in finer detail.
"Minor issues": we will improve readability of sections 4 and 5 and simplify notation as suggested.