Thank you very much for your reviews.

Reviewer 1:

1. Regarding the parameters in Section 4.2, I ran my experiments with 2 more sets of parameters, where the target dimensions are much lower than in Figure 6. The trends match trends in the submission as expected.

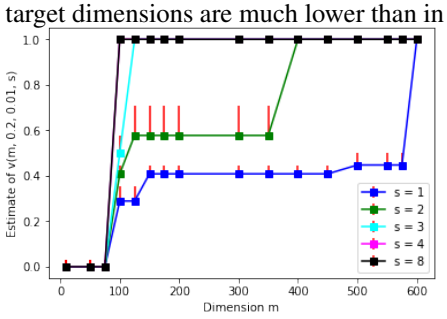

Figure 1: Phase transitions of $\hat{v}(m, 0.2, 0.01, s)$

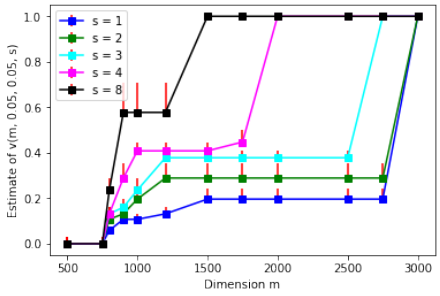

Figure 2: Phase transitions of $\hat{v}(m, 0.05, 0.05, s)$

2. Regarding experimental design in Section 4.2, recall my goal was to compute an estimate $\hat{v}(m, \epsilon, \delta, s)$ of $v(m, \epsilon, \delta, s)$. As mentioned in footnote 20, the design is based on Section 3.1 of [13] (for $s = 1$).

   Here is a more detailed explanation of how $\hat{v}(m, \epsilon, \delta, s)$ is computed: I test using a set W of values w spaced between 0.03 and 1 (see footnote 19) . For *each* $w \in W$, I compute $\hat{\delta}(s, m, \epsilon, w)$, an estimate of failure probability on the specific binary vector x^w where the first $1/w^2$ entries are nonzero. Then, I let $\hat{v}(m, \epsilon, \delta, s)$ be the max. value in W such that $\hat{\delta}(s, m, \epsilon, w) \leq \delta$ for all $w \in W$ where $w \leq \hat{v}(m, \epsilon, \delta, s)$.

   How do I compute $\hat{\delta}(s, m, \epsilon, w)$*?* As mentioned in the submission, I estimate $\mathbb{P}_{A \in \mathcal{A}_{s,m,n}}[\|Ax^w\|_2 \notin (1 \pm \epsilon) \|x^w\|_2]$ by computing the projected norm for $T = 100,000$ samples of a block sparse JL matrix.

   Why does it suffice to only consider sparse vectors x^w*, rather than all vectors in* S_v*?* As mentioned in footnote 21, I show in the proof of Theorem 1.5 that asymptotically, if a "violating" vector (i.e. x s.t. $\mathbb{P}_{A \in \mathcal{A}_{s,m,n}}[\|Ax\|_2 \notin (1 \pm \epsilon) \|x\|_2] > \delta$) exists in S_v, then there's a "violating" vector x^w for some $w \leq \Theta(v)$. Thus, the estimate $\hat{v}(m, \epsilon, \delta, s)$ will approach $v(m, \epsilon, \delta, s)$ up to constants as $T \to \infty$ and as precision in W goes to ∞ (if ϵ, δ are sufficiently small for the "violating" vector asymptotics to kick in).

3. Regarding the Section 4.1 experiment, the failure probability actually increases to a local maximum somewhere in $12 \leq s \leq 16$, and then decreases when $s \geq 16$, reaching lower than the value at $s = 8$ by the time $s = 20$. When $\epsilon = 0.07$ and $m = 500$, there is similarly a local maximum (somewhere in $24 \leq s \leq 32$) followed by a decrease. The phenomenon of non-monotonicity in s can also be observed on synthetic data in Figure 6 in the submission: for example, when $\delta = 0.05$, $\epsilon = 0.02$, $m = 12000$, we see that $\hat{v}(m, \epsilon, \delta, 4) < \hat{v}(m, \epsilon, \delta, 3)$. I'd like to emphasize that my asymptotic theoretical results characterize the macroscopic behavior of $v(m, \epsilon, \delta, s)$, and do not preclude the existence of constant factor fluctuations for small changes in parameters.

4. Regarding [29], I will certainly add this reference – thanks for pointing this out. I agree that this reduction gives lower bounds for a distribution similar to the uniform sparse JL distribution. However, as you mentioned, the lower bounds differ from my work in the following ways: (a) they do not match the bounds in Theorem 1.5, since they would not recover the branch 3, and (b) the distribution resulting from this reduction gives s nonzero entries that are independently selected (potentially resulting in a multiset), which is different than the uniform sparse JL distribution. Regarding (b), in fact, Theorem 5.1 in the JACM version of [18] shows that this "multiple hashing" distribution requires an extra (roughly) $\log(1/\delta)$ factor on the sparsity to satisfy (1).

5. Regarding obtaining a fine-grained understanding of each vector in $\mathbb{R}^n$, I agree this would be an interesting result, though it would not immediately follow from the techniques that I present in this submission.

Reviewer 3:

1. I appreciate that you read through my supplementary material, and I will certainly address the typos you noted. (Regarding the specific typo on p. 12, it should say $T \geq \max(\frac{se}{mv^2}, 3), T \leq \log(Tmv^2/s)$.)

2. Regarding your comment about feature vectors, I agree that considering vectors with restricted ℓ_∞-to-ℓ_2 norm ratio is also interesting in its own right from a theoretical perspective. This perspective also does nicely motivate my lower bound on dimension-sparsity tradeoffs (footnote 3, Corollary 1.3 in the supp. material).

   Nonetheless, my experiments in Section 4.1 link my analysis to feature vectors by considering the performance of sparse JL on feature vectors. I specifically evaluate the performance of sparse JL on bag-of-words feature vectors in two real-world datasets: News20 and Enron emails. Note that [29] also evaluates on bag-of-words datasets (in collaborative spam filtering, and the work initiates the study of vectors with restricted ℓ_∞-to-ℓ_2 norm ratio in this context). Subsequent work [13, 10] also experimentally considers the performance of feature hashing on bag-of-words feature vectors from News20 and a collection of NeurIPS papers.