We thank the area chair and the four reviewers for their careful reading and helpful comments. We will begin with some general clarifications and then follow with specific response line by line.

Our Contributions For the non-convex problem of matrix sensing, we define a function $\delta_{\rm soc}(X)$ that gives a *precise* threshold on the number of samples need to prevent X from becoming a spurious local minima. Although $\delta_{\rm soc}$ is difficult to compute exactly, we obtain a *closed-form*, sharp lower bound using convex optimization. As a result, we are able to characterize the *tradeoff* between the quality of the initial point and the sample complexity.

Comparison with previous results on local convergence: Various previous works have shown that linear convergence occurs around a small, fixed neighborhood of the global min (see Bhojanapalli et al., Tu et al., etc). The proof techniques are similar: restricted local convexity holds when the sample size is sufficiently large. However, these proof techniques are incapable of charactering how the optimization landscape changes as sample complexity increases. Our work paints the full picture: the problem becomes more 'non-convex' (requiring more samples to eliminate spurious local min) as we get further and further away from the global min. Once outside $\mathcal{B}_\varepsilon$, it becomes *necessary* to rely on the global guarantees of Bhojanapalli et al. In contrast, previous work on local convergence only show convexity in a small neighborhood, and tells us nothing about the landscape outside that small neighborhood.

How to find an initial point: As reviewer 1 points out, the main concern of our paper is understanding how the landscape changes with sample complexity. Therefore, we chose to view the initial point as a part of the *problem structure*. Nevertheless, there is a substantial body of previous work (e.g. Bhojopanalli et al., Tu et al., Candes et al.) that separately studies the problem of finding a good initialization. One possible difficulty, as reviewer 4 notes, is that some of these methods, such as spectral initialization, already require a large sample size. *But we emphasize that this is not the only way to get an initial point.* For example, matrix sensing arises in the electric grid application under the name "state estimation". Here, the ground truth corresponds to a physical quantity of interest. Domain-specific heuristics that depend on physical and engineering intuition are able to deliver high quality initial points that are then further refined via non-convex optimization.

Response to reviewer 1: We thank the reviewer for the nice summary of our paper. We will move the related works section towards the end of the paper. Regarding the second question in section 3, we note that GD will always stay in the ε-ball when the sample size is large (but still on the order of $O(nr)$). In this case the inner product between $\nabla f(X)$ and $\nabla \|XX^T - ZZ^T\|_F^2$ is always positive. When the sample size is smaller, we can rely on problem structures to prevent the algorithm from leaving the neighborhood. For instance, with any descent algorithm, we are guaranteed to stay in the region if we initialize within a smaller interior (See [23]).

Response to reviewer 2: We thank the reviewer for the positive feedback. We agree that the title of the paper is indeed too general and we will change it to *How Many Samples is a Good Initial Point Worth in Low-Rank Matrix Recovery?*

Response to reviewer 3: We thank the reviewer for very detailed comments. (1) We agree that more motivation should be provided for the matrix sensing problem. We have added a brief section in the intro that discusses the application of matrix sensing in problems like quantum state tomography, metric learning, and electric grids. We also clarified our assumptions: the measurement matrices $A_1, \dots, A_n$ are fixed, and can be from *any* RIP ensemble. (2) Regarding the tightness of our lower bound: the plot in figure 1 shows the rank-1 case, where the bound has been shown to be tight for all ε (See [23]). In the high-rank case, $\delta_{\rm foc}$ is very close to 1 when ε is small, as indicated by Theorem 5. Since $\delta_{\rm foc} \le \delta_{\rm soc} < 1$, the gap between $\delta_{\rm foc}$ and $\delta_{\rm soc}$ is small. When ε becomes large, we switch to the *global* lower bound $\delta_{\rm soc}(\mathbb{R}^{n\times r}) = 1/5$, which is again exactly tight. (3) Arguably, matrix sensing is one of the handful non-convex problems that admits rigorous theoretical analysis, and our work provides deeper understanding of how non-convexity can be overcome with more training samples. We believe this is an important step towards understanding the relationship between sample complexity and the optimization landscape in deeper models. (4) Notice that when the number of measurements is below the threshold defined by $\delta_{\rm soc}$, our results guarantee that there *exists* some choice of the measurement ensemble $\mathcal{A}$ such that the problem will have a spurious local minima. However, sampling from sub-Gaussians distributions in general does not find these adversarial cases. This is indeed a subtle point, and we have added a brief discussion in the numerical results section.

Response to reviewer 4: We thank the reviewer for the helpful comments. For the concerns raised in section 3, please refer to our discussion at the beginning. We emphasize that our main contribution is *not* improved RIP-conditions. Rather, it is a new proof technique that establishes a *tradeoff* between sample complexity and the quality of the initial point. This is something that previous methods based on local convexity are incapable of characterizing, since their analysis depends on a *fixed* neighborhood. Note that lemma 4.2 in [1] only bounds the distance in the *subspace* spanned by the column of U, and the error along the orthogonal direction can still be large. Therefore, this lemma can't actually eliminate spurious critical points, even when δ is arbitrarily small. In contrast, our analysis finds the precise number of samples to prevent *any* point from becoming a spurious critical point, allowing us to describe how the optimization landscape 'evolves' as sample complexity increases.