We would like to thank all three reviewers for their constructive comments and suggestions which will help us improve the quality of this paper (ID: 1916).

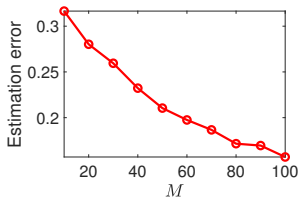

Fig.1: estimation error bound.

Reviewer $\#1$**:**

Q1: I would expect more explanations of the main result, e.g., a proof sketch with some intuition about the significance of this result.
A1: Thank you for pointing this out. We will add a new paragraph or subsection after we present Theorem 2.1 to include a proof sketch with some intuition about the significance of our result. In this section, we will also include the estimation error bounds obtained from Theorem 2.1 and some discussion of the parameters ϵ and η as requested by reviewers $\#2$ and $\#3$.

Q2: In the Applications section, constants can be hidden. I would suggest the authors hide details like the division of regions and formulas in line 196 to make more space for the main result section.
A2: We would like to emphasize that the constants and formulas in line 196 are critical. These constants are carefully chosen to argue that the two applications satisfy the requisite assumptions. Similarly, we feel the partition regions are also very important, and we did receive positive comments from reviewers $\#2$ and $\#3$ concerning the paper organization. So, we are hesitant to remove these details. However, we will be allowed a ninth content page if our paper is accepted, and we plan to devote this page to extending the discussion in the main result section. If that turns out to be insufficient, we plan to reduce the lemmas in the application section into two informal lemmas, and move the current formal lemmas to the supplementary material.

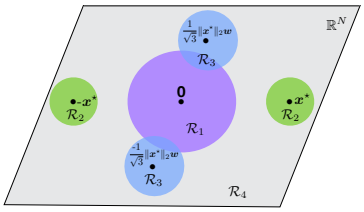

Fig.2: partition regions in phase retrieval.

Q3: Do experiments on more general settings for the first application.
A3: In the revised version, we will incorporate Fig.1, which shows the distance between the local minima of population and empirical risk in the problem of matrix sensing with $k = 2$, $r = 3$, and $N = 8$.

Q4: Have a figure with partition regions for phase retrieval. Use a figure to illustrate the assumptions or results.
A4: We will add Figs. 2 and 3 in the revised version. Fig. 3 is used to illustrate the assumptions in the example of phase retrieval. In particular, we set the parameters in Example 1.2 as $N = 1$, $x^\star = 1$, and $M = 30$. We display the population risk and empirical risk together with their gradients and Hessians in Fig. 3. One can see that in the small gradient region (the three parts between the light blue vertical dashed line), $|\lambda_{\min}(\text{hess } g(\mathbf{U}))|$ (which equals the absolute value of the Hessian since $N = 1$) is bounded away from zero. With enough measurements, e.g., $M = 30$, the gradients and Hessians of the empirical and population risk are close to each other. In addition, we think our simulation figures are good illustrations of the main results in Theorem 2.1, i.e., the relationship between critical points of empirical and population risk. We will add more details on this in the main result section.

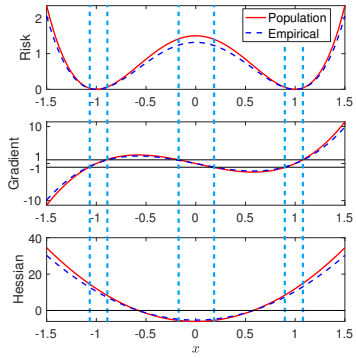

Fig.3: phase retrieval with $N = 1$.

Reviewer $\#2$**:**

We appreciate the positive comments and will add a detailed discussion on how to obtain estimation error bounds.

Reviewer $\#3$**:**

Q1: The main assumption in the paper does not include degenerate minima, which is surprising (a drawback).
A1: A degenerate local minimum of the population risk can correspond to a strict saddle point of the empirical risk due to its randomness. Thus, we do not believe there exists a correspondence in the case of degenerate local minima.

Q2: Add an intuitive description of the meaning of ϵ and η in the examples of matrix sensing and phase retrieval and an explanation of the proof strategy for obtaining the parameters. Is there a standard way to compute ϵ and η?
A2: For phase retrieval, note that $|\lambda_{\min}(\nabla^2 g(\boldsymbol{x}))|$ and $\|\nabla g(\boldsymbol{x})\|_2$ roughly scale with $\|\boldsymbol{x}^\star\|_2^2$ and $\|\boldsymbol{x}^\star\|_2^3$ in the regions near critical points, which implies that η and ϵ should also scale with $\|\boldsymbol{x}^\star\|_2^2$ and $\|\boldsymbol{x}^\star\|_2^3$, respectively. For matrix sensing, in a similar way, $|\lambda_{\min}(\text{hess } g(\mathbf{U}))|$ and $\|\text{grad } g(\mathbf{U})\|_F$ roughly scale with λ_k and $\lambda_k^{1.5}$ in the regions near critical points, which implies that η and ϵ should also scale with λ_k and $\lambda_k^{1.5}$, respectively. We will incorporate a more detailed discussion on this in the revised version. To obtain these parameters, one can lower bound $|\lambda_{\min}(\text{hess}(g))|$ in a small gradient region. In this way, one can adjust the size of the small gradient region to get ϵ, and use the lower bound for $|\lambda_{\min}(\text{hess}(g))|$ as η. In the case when it is not easy to directly bound $|\lambda_{\min}(\text{hess}(g))|$ in a small gradient region, one can also first choose a region for which it is easy to find the lower bound, and then argue that the gradient has a large norm outside of this region, as we did in this paper.