Reviewer 1. We appreciate your invaluable comments and questions. We address your concerns below.

R1.1 **Huber loss vs. ℓ^1-loss.** Our choice of Huber loss rather than ℓ^1-loss is to simplify theoretical analysis. Undoubtedly, ℓ^1-loss is a more natural sparsity promoting function and performs better than Huber. We will refine our statement in the revision to make this fact more clear and transparent. When ℓ^1-loss is utilized, the experiments tend to suggest that the underlying kernel and signals can be exactly recovered even without LP rounding; see the figure below in which rounding is not applied to ℓ^1-loss. However, on theory side, the subgradient of ℓ^1-loss is *non-Lipschitz* which introduces tremendous difficulty in controlling suprema of random process and perturbation analysis for preconditioning for our problem. We leave analyzing ℓ^1-loss for future research which we will discuss in the final draft.

R1.2 **Relation to cross-relation methods.** We apologize for omitting this literature (e.g., G. Xu et al. TSP'95, Y. Lin et al. NeurIPS'07, K. Lee et al. SIIMS'18), which consider similar problems that we will include in revision. In comparison, *(i)* Lin et al. proposed an ℓ^1-regularized least-squares method based on convex relaxation. The convex method could suffer similar sparsity limitation as [19], and it limits to 2 channels without theoretical guarantees; *(ii)* Lee et al. proposed an eigen approach for FIR signal model. Their provable efficient method considered short signal $\mathbf{x}_i$ which lives in given random subspaces, thus it cannot directly handle our case with random sparse nonzero support.

R1.3 **Clarification of the set $\mathcal{S}_\xi^{i\pm}$.** Our original description that the union of $\mathcal{S}_\xi^{i\pm}$ gives a full partition of $\mathbb{S}^{n-1}$ is inaccurate. Indeed, we consider these sets because their union excludes all saddle points, but covers a large portion of measure over the sphere for small ξ. For each individual set, we will make the inclusion $\mathbf{q} = \mathbf{e}_i$ clear and well-defined.

R1.4 **Closeness of $\mathbf{RQ}^{-1}$ and I.** We will include the precise approximation error from Lemma H.4 of supplementary.

R1.5 **Algorithmic implementation details.** Experimentally we use linesearch for both the stepsizes in RGD and LP rounding for optimizing all losses. In revision, we will expand Line 230 with more details and release code.

In revision, we will add a conclusion and address other minor issues without detailed explanation due to space limit.

Reviewer 2. We sincerely thank you for your appreciation of our work. For reproducible research, we will release well documented code of this work on Github, and correspondingly provide a link in the revised draft.

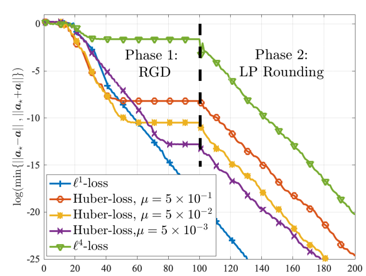
R2.1 **Smoothing parameter μ.** Here, the parameter μ in Huber introduces a *tradeoff* between sample complexity and recovery accuracy. As shown in Theorem 3.1, the sample complexity p depends *inversely* on μ: larger p is required for smaller μ, and vice versa; on the other hand, Figure 2 (or the revised figure below, as suggested) shows smaller μ produces higher recovery accuracy in Phase 1. We will discuss this around Theorem 3.1 and experiment section.

Reviewer 3. We really appreciate your constructive criticism and valuable feedbacks, that we address as follows.

R3.1 **Sample Complexity.** We agree there is a large sample complexity *gap* between our theory and practice. From the degree of freedom perspective (e.g., Eric Moulines et al. TSP'95), a constant p is also seemingly enough in our case. However, as the problem is highly nonconvex with unknown nonzero supports of $\mathbf{x}_i$s, to have provable efficient methods, we conjecture that paying extra log factors $p \geq \Omega(\mathrm{poly}\log(n))$ is necessary as stated in Line 56, 139 and 251, which is empirically confirmed by Figure 4. This is similar to recent provable efficient method on FIR model (K. Lee et al. SIIMS'18). On the other hand, we believe our *far from tight* sample complexity $p \geq \Omega(\mathrm{poly}(n))$ is due to the looseness in our analysis: *(i)* loose control of summations of dependent random variables, and *(ii)* tiny gradient near the set boundary for concentration. We will discuss this in revision and leave improvement for future work.

R3.2 **Clarification of rounding with unknown rotations.** We do *not* need to know $\mathbf{Q}$ for solving LP rounding. Footnote 9 and Appendix I provide more details of the *actual* problem form we are solving. The reason we stated LP rounding in the rotated space as (14) (typo: $\mathbf{u}$ should be $\mathbf{q}$ in (14)) is *only* for the convenience of introducing subsequent results. Recall the deduction from (4) to (9), we can *reversely* get back the actual form in Footnote 9 by plugging $\mathbf{q} = \mathbf{Q}\mathbf{q}'$ into (14) (with an abuse of notations of $\mathbf{q}$ and $\mathbf{q}'$), where $\bar{\mathbf{r}}$ is the *actual solution* of optimizing (4) in Phase 1. Therefore, $\mathbf{Q}$ is *not* needed. We will make this involved narrative more clear in revision.

R3.3 **Simulation in Figure 2.** Following reviewer's suggestion, in the right figure we show the convergence in a progressive way for optimizing ℓ^4 and Huber losses, with the same setup as in Figure 2. We observe that (i) in Phase 1, the reconstruction errors stagnate for both ℓ^4 and Huber losses before rounding is applied, and (ii) in Phase 2, the projected subgradient method for LP rounding converges *linearly* to a target solution. For the ℓ^1-loss, it seems that rounding is *not* necessary. Per our R1.1 to Reviewer 1, analyzing this behavior is the subject of future work.



R3.4 **Other technical issues.** We briefly address other minor technical issues as follows: *(i)* Equation (2) is to show *intrinsic* shift-scaling symmetry, so that we can only solve to a shift equivalence; *(ii)* the DFT matrix $\mathbf{F}$ is unnormalized, as shown on Line 389 of the supplementary; *(iii)* Figure 1 plots the function values over the sphere, where cooler color denotes smaller values, and vice versa. The target solutions (red dots) stays much closer to global minima for Huber and ℓ^1 losses than ℓ^4-loss; *(iv)* the problem becomes trivial when $\theta \leq 1/n$ because $\theta n = 1$ so that $\mathbf{x}_i$ tends to be a δ-function; *(v)* We will mention the result by Cosse. As it is a convex method following [19], it may suffer similar limitations.

Due to space limit, for other questions we refer Reviewer 3 to our response to Reviewer 1 (e.g., R1.1, R1.2, and R1.3).