We would like to thank the reviewers for the careful and thorough reading of our submission. We are appreciative of the many suggestions for improvements and insightful questions. In the limited space below, we respond to some of the main concerns raised by the three reviewers.

The novelty of our analysis in the case $p = 2$ [Reviewer 1 and Reviewer 2] As the reviewers note, for the case $p = 2$, the tight approximation bound of CSS is already known in the classic work [14]. However, our work has a different goal: to provide a unified way of analyzing approximation bounds for all different values of p. With this goal in mind, we introduced the Riesz-Thorin theorem as a general framework, and found that the existing analysis in [14] unfortunately does not fit in this framework, due to the technical differences stated below.

The main technical difference can be found in line 219-222: the Riesz-Thorin theorem requires that Equation (4) should hold for *all* $\{a_t\}_{t\in\binom{[m]}{1}} \in \mathbb{C}^{\binom{[m]}{1}}, \{b_J\}_{J\in\binom{[m]}{k}} \in \mathbb{C}^{\binom{[m]}{k}}$. In comparison, to prove that CSS is a $\sqrt{k+1}$ approximation, one only needs to show that the equation holds for $\{a_t\}_{t\in\binom{[m]}{1}} \in \mathbb{C}^{\binom{[m]}{1}}, \{b_J\}_{J\in\binom{[m]}{k}} \in K$, where K is a subset of $\mathbb{C}^{\binom{[m]}{k}}$ defined as $K = \{\{b_J\}_{J\in\binom{[m]}{k}} | b_J = \det(S_J), S \in \mathbb{C}^{k\times m}\}$. It is easy to see that the requirement of Riesz-Thorin is significantly stronger, since the set K is determined by only km parameters (the matrix S), while $\mathbb{C}^{\binom{[m]}{k}}$ is a $\binom{m}{k}$-dimensional space. In other words, we need a much stronger inequality in order to apply Riesz-Thorin theorem, hence we provided a brand new proof for the $p = 2$ setting (line 421-423, as mentioned by reviewer 1) which is completely different from [14].

The choice of p in l_p low rank approximation [Reviewer 1 and Reviewer 2] While our paper is mostly theoretical, we do believe that choosing an appropriate value of p can make a difference in practice. The ℓ_p low rank approximation problem has attracted interest relatively recently, see for instance ICML 2017 paper [13]. A related problem ℓ_p linear regression however has been studied extensively in the statistics community, and these two problems share similar motivation. In particular, if we assume a statistical model $A_{ij} = A^\star_{ij} + \varepsilon_{ij}$, where $A^\star$ is a low rank matrix and ε_{ij} are i.i.d. noise, the different values of p correspond to the MLE of different noise distributions, say $p = 1$ for Laplacian noise and $p = 2$ for Gaussian noise. To capture a broader range of realistic noises in complex datasets, it is beneficial to expand our choices beyond the standard ones $(1, 2, \infty)$.

Sampling Based Algorithms [Reviewer 2] This is definitely an important direction for the future works that we intend to explore. Our determinantal weights are indeed inspired by [14] and we will include the comparison in our final version. The sampling methods for the $p = 2$ case, e.g. volume sampling, are closely related to the determinantal point process. To generalize this approach to ℓ_p setting, we need an efficient way to implement the exponentiated variant of volume sampling (i.e. sample a subset S with probability proportional to $\det(V_S)^\alpha$). To the best of our knowledge, this problem has not been resolved yet. In fact, even computing the normalizing constant of the distribution is difficult - it was stated as an open problem in Section 7.2 of the survey "Determinantal point processes for machine learning". We refer the reviewer to the NeurIPS 2018 paper "Exponentiated Strongly Rayleigh Distributions" for recent progress on this problem.

Clarification on Theorem 1.2 [Reviewer 3] We thank the reviewer for pointing out the ambiguity of the informal statement of the theorem. We proved that there are infinitely many different values of k, such that for each k, there exists a matrix A such that CSS cannot do better than $(k+1)^{1-\frac{1}{p}}$ approximation. The dimensions of the matrix A (m and n) are not fixed for these different k's.

Optimization landscape of low rank approximation [Reviewer 3] We refer the reviewer to the paper "Neural networks and principal component analysis: Learning from examples without local minima" where the paper shows that for low-rank approximation problem even as easy as PCA ($p = 2$), saddle points exist. We are happy to modify the relevant sentence in our paper as "Unfortunately the loss surface of the problem suffers from many saddle points" and replace [12] with the reference above.

The optimality gap of the bound when $1 < p < 2$ [Reviewer 3] (line 83) We note that the lower bound (Theorem 1.2) also applies to the case $1 < p < 2$. Therefore, we have an upper bound of $(k+1)^{\frac{1}{p}}$ and a lower bound $(k+1)^{1-\frac{1}{p}}$, hence we were correct on the $(k+1)^{\frac{2}{p}-1}$ optimality gap.

Improvements on Exposition [Reviewer 1,2,3] We thank all three reviewers for their suggestions on the exposition and will take them into account in the final version. In particular, we will clarify the difference of ordered and unordered sets; add more explanation to the reduction in the proof of Lemma 2.2; improve the sentence structure in Lemma 2.3 and include the definition of CUR factorization in the introduction.