We are thankful to the reviewers for their insightful comments and suggestions. Below we address each reviewer's questions and concerns, with the questions and concerns briefly rephrased in blue.

Response to Reviewer #2.

[Choice of r]: r is $\Theta(\log n)$. Actually, we do not really need to choose r. In Algorithm 1, we drop $1/20$ fraction of T_{i-1} to get T_i in the i-th iteration of the outer loop. Thus, the outer loop will have at most $O(\log n)$ iterations. We use r to simplify the notations in our analysis.

[Use of Cramer's rule]: Consider a rank k matrix $M \in \mathbb{R}^{n\times(k+1)}$. Let $P \subseteq [k+1], Q \subseteq [n], |P| = |Q| = k$ be such that $|\det(M_P^Q)|$ is maximized. Since M has rank k, we know $\det(M_P^Q) \neq 0$ and thus the columns of M_P are independent. Let $i \in [k+1] \setminus P$. Then the linear equation $M_P x = M_i$ is feasible and there is a unique solution x. Furthermore, by Cramer's rule $x_j = \frac{\det(M^Q_{[k+1]\setminus\{j\}})}{\det(M_P^Q)}$. Since $|\det(M_P^Q)| \geq |\det(M^Q_{[k+1]\setminus\{j\}})|$, we have $\|x\|_\infty \leq 1$.

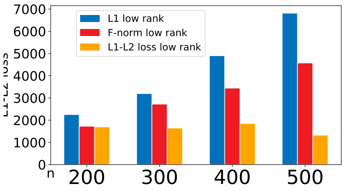
We just realized that there is a typo in the equation of line 232, and we correct it as the following:

$$R_{M^*}(H) = \arg\max_{P:P\subseteq H} \left\{ \left|\det\left((M^*)_P^Q\right)\right| \;\middle|\; |P| = |Q| = \operatorname{rank}(M^*_H), Q \subseteq [n] \right\}.$$

["Necessary" conditions and zero-one law]: We can change the wording, and indeed what we meant is that if we're missing any one of the mentioned properties, then we can find an example function for which there is no good column subset selection. On the other hand, if we have all the properties, we have a good column subset selection. This is what we meant in the title by zero-one law, though we are happy to change the wording. We believe this is sometimes loosely what is intended by the word characterization, as also stated by reviewer 4 - that "it needs to satisfy approximate triangle inequality, and monotone property", otherwise there are counterexamples. But, again, we would change the wording.

Response to Reviewer #3.

[Experiments on other loss functions]: We are glad to report results with other loss functions in the final version. Due to the page limit of the response, here we only present the result with the loss function "$\ell_1 - \ell_2$" (i.e., $2(\sqrt{1 + x^2/2} - 1)$). The setting is the same as the experiments represented by Figure 2 except that the loss function used is now "$\ell_1 - \ell_2$".



[Minor question, choice of r]: See the response to Reviewer #2.

[Textual clarity]: We thank the reviewer for the comments on the presentation.

Response to Reviewer #4.

[Lower bound]: We discuss the necessity of the triangle inequality and monotone property in Appendix B. In Appendix B.1, we show that there is no good column subset selection for the jumping function which is monotone but does not satisfy the approximate triangle inequality. In Appendix B.2, we show that there is no good column subset selection for the ReLU function which has the triangle inequality but does not satisfy approximate monotonicity. We can put both results into the main body in the camera ready version.

[Presentation]: We thank the reviewer for the comments on the presentation. We will fix these issues in the camera ready version.

- In the equation between line 232 and line 233, the max is over all possible choices of P and Q, and $R_{M^*}(H)$ only takes the value of the corresponding P.
- Lines 158 to 160: if we first choose a random subset T of k columns and then randomly choose another column i, then $T \cup \{i\}$ is a random set of $k+1$ columns. Consider the submatrix $A_{T\cup\{i\}}$. If we want to choose a subset of k columns from $A_{T\cup\{i\}}$ to approximate $A_{T\cup\{i\}}$, then by symmetry, with probability $1/(k+1)$, i may not be in the optimal selection of k columns. Furthermore, since $T \cup \{i\}$ contains $k+1$ random columns, the expectation of the best rank-k approximation to $A_{T\cup\{i\}}$ is at most a $(k+1)/n$ fraction of the optimal rank-k approximation cost for A. By Markov's inequality, we obtain Equation (1).
- Lines 168 to 179: at a high level, in the analysis of [62], the authors need to conceptually normalize the columns to unit norm. However when the loss function is not scale-invariant, their algorithm is not able to normalize and thus the previous analysis completely breaks, as we explain starting at line 180. This is why we need to develop new techniques for analyzing our algorithm.