Response to Reviewer #1

[Discrepancy between the algorithm analyzed and the evaluated]: Our theory motivated the algorithm we implemented, though we tweaked it empirically and observed improvements by taking the median. Indeed, our theory shows that aggregating multiple columns helps reduce the cost for the ℓ_1-loss in our distributional setting, which is known to be impossible in a worst-case setting (see below). We initially implemented the average, and that performed well as predicted by our theory, but taking the median performed even better on our datasets. However we agree that the latter does not have theoretical guarantees, and so we are also happy and will include figures for taking the average.

[Order of the polynomial in line 64-65]: The exponent depends on the exponent of the polynomial of the ℓ_1-regression solver used. If ℓ_1-regression solvers improve, then the running time of our algorithm also improves. Furthermore, the order of the polynomial also depends on the constant p. For larger p, the order of the polynomial is smaller. We did not explicitly compute the order, but we stress that before our work no polynomial of any order was known.

[Apply Algorithm 2 for best rank-k approximation]: We do not know if the median algorithm can be analyzed for general k to give a theoretical result for low rank approximation with the ℓ_1-loss, and leave this as an intriguing open question, inspired by our experiments.

Response to Reviewer #2

[Median heuristic]: See response to Reviewer #1 above.

[Clarity of Subsection 1.2 and Subsection 2.2]: At the end of page 3, we should mention that the cardinality of a random set is at most $\mathrm{poly}(k/\varepsilon)$ and there are $\mathrm{poly}(k/\varepsilon)$ of them. In Subsection 2.2, the use of Cramer's rule is as follows. 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 $x_j = \det(M^Q_{[k+1]\setminus\{j\}})/\det(M_P^Q)$. Since $|\det(M_P^Q)| \geq |\det(M^Q_{[k+1]\setminus\{j\}})|$, we have $\|x\|_\infty \leq 1$.

[Applicability/significance of the method beyond the ℓ_1-loss]: Our averaging idea is very general and we expect it to be useful for other loss functions, which we leave as an open question. We focused on ℓ_1 since there is a known lower bound for column subset selection with ℓ_1-loss (see [24], discussed more below), which we bypass in our setting.

Response to Reviewer #3

[Overall comparison and comparing with the bounds obtained when ℓ_1-loss is used in #3303]: The other paper concerns characterizing when one can obtain good low rank approximations for arbitrary loss functions and is for worst-case inputs, while for the ℓ_1-loss it is known [24] that it is impossible to obtain subsets of fewer than $\mathrm{poly}(k)$ columns spanning any $\sqrt{k}$-approximate low rank approximation. Here, under our distributional assumptions we are able to get very accurate, $(1+\varepsilon)$-approximate low rank approximations for ℓ_1 with only $O(k \log n) + \mathrm{poly}(k/\varepsilon)$ columns. Further, we show our distributional assumption is necessary. Thus, the other paper #3303 cannot be used to obtain a $(1+\varepsilon)$-approximation since it is simply not possible without distributional assumptions – indeed, as the reviewer states, that paper obtains an $O(k \log k)$ approximation. Further, showing that our distributional assumptions suffice is quite involved. We believe when high accuracy is required, this paper should be used, while when one wants an arbitrary loss function, which might not be scale-invariant (so well beyond ℓ_p norms, e.g., Huber loss), then one should use that paper.

[When necessary conditions are met for both papers]:The tradeoff is not only between time and approximation but also between the number of columns selected and the approximation. #3303 selects $O(k \log n)$ columns but only achieves $O(k \log k)$ approximation while this paper selects $O(k \log n) + \mathrm{poly}(k/\varepsilon)$ columns and achieves $(1+\varepsilon)$ approximation. Technically, this paper mainly focuses on showing how an extra $\mathrm{poly}(k/\varepsilon)$-sized sample of columns can help reduce the error to $(1+\varepsilon)$ while #3303 focuses on why a recursive filtering approach works for general loss functions.

[Mean and standard deviation]: We repeated each experiment 25 times. In all experiment settings, our average approximation ratio is the best among compared algorithms. In addition, the standard deviation is also small. For example, for mfeat dataset, $k = 1$ and 1.1-stable random noise, the mean approximation ratio of SVD, L1LowRank, Uniform and Ours are $3.05, 4.12, 1.27, 1.01$ respectively, and the standard deviations are $0.47, 1.46, 0.09, 0.001$ respectively. The randomness is over both the data and the algorithm. Due to the page limit of the response, we will report all means and standard deviations in the final version.

Response to Reviewer #4

[Success probability]: Though the success probability is with respect to both the data and the randomness of the algorithm, all stated .999 probabilities can be made $1 - n^{-\Omega(1)}$. In Lemma 2.3, the size of H (the set of columns with large entries) slightly changes, and this changes line 297, where $T' \leq O(\cdots + |H|)$. Notice that in line 290, qt is $n^{o(1)}$, so no change is needed. In Lemma 2.3, we can make $|H| \leq n^{1-(p-1)/2+\delta}$ for some sufficiently small positive constant $\delta \leq (p-1)/2$ (e.g., $\delta = (p-1)/4$), and still use Markov's inequality. This is desired since what we want is to make $|H| \leq n^{1-\Omega(1)}$. Furthermore, this will happen with probability at least $1 - n^{-\delta} = 1 - n^{-\Omega(1)}$.

[Detailed question]: 1. The max is over P and Q while $R_{A^*}(S)$ only takes the value of the corresponding P. 2. The reason that we spent considerable effort is that our moment condition seems to be a considerably weaker assumption than typical assumptions, e.g., it is much weaker than assuming a bounded variance, and thus while it resembles central limit theorems, it is definitely different as the random variables we consider may not even have a variance.