We thank the reviewers for the valuable time they have invested during this difficult period to review the paper and provide helpful suggestions for improving the manuscript. We also appreciate their complimentary comments and praise for the paper's theoretical contributions. In the following we address their comments, suggestions and questions.

Reviewers 1 and 2 ask about the precise relation between the parameter matrix M and the observation matrix X, expressed by the notation $X \sim M$. As described in line 62 of section 1.2, the notation means that M generates X via one of the five models described in section 1.1. Later in the paper, we use this notation without reminding the reader of its meaning. Let us further clarify this notation here by briefly demonstrating it for two of the five models. In the paper's final version we similarly clarify the precise meaning of $X \sim M$, and repeat it before Theorem 2.

In the first model, *distribution matrices*, M is a distribution matrix over $[k] \times [k]$ scaled by the number of times the distribution is sampled. Each time the distribution is sampled, a single independent value (i, j) is observed, and X is the total number of times element $(i, j) \in [k] \times [k]$ was observed among all samples taken. In the *Bernoulli matrices* model, M is a matrix of Bernoulli parameters, and $X_{i,j} \sim \text{Ber}(M_{i,j})$, independently, hence each $X_{i,j}$ is either 0 or 1.

The paper's unified formulation (lines 53-57) implies that in all 5 models (slight adjustment for #5), $X_{i,j}$ is the number of times (i, j) is observed, and when $X \sim M$, $E[X] = M$. Hence the expected number of samples is $E[||X||_1] = ||M||_1$.

Reviewer 2 asks whether the paper's title is justified as we learn just the means of $X_{i,j}$, not their distribution. As we hope is now clear, we learn the parameter matrix M that in all cases (except a small modification for collaborative filtering) determines the whole distribution of X. The fact that $E(X) = M$ is just a useful byproduct of our unified framework.

Reviewer 3 asks if multiple independent draws (of X) from the matrix M may improve its estimate. That is correct. However, please note that such re-sampling is already accounted for in our models. For example, if t i.i.d. instances $X^1, \dots, X^t \sim M$ are generated using the Bernoulli model, the sufficient statistics would be the sum $X = \sum_i X^i$. This correspond exactly to $X \sim M' = tM$ using the Binomial model. A similar relation holds for all other models.

All four reviewers acknowledge the paper's technical contributions. However three reviewers point out that the algorithm is simple enough to implement and the experiments will improve the paper. Similar to the prior work that this paper improves on, e.g., [HKKV18] and [BCLS17], we viewed it mostly as a theoretical contribution, hence did not demonstrate its empirical efficacy. However, the algorithm is indeed simple to implement, and given the reviewers encouraging interest, we will gladly include simulations in the final version. From preliminary simulations, we expect good performance.

Reviewer 2 writes that the lower bound (Theorem 1) is well known, and should not be mentioned as a contribution. We are not aware however of such a lower bound for low-rank matrices, and if one is pointed out to us, we would gladly cite it. Yet the reviewer has a point in that the lower bound is very simple. We therefore don't even "prove" it, we just informally describe (lines 88-91) how it follows from previous work. We mention the lower bound to provide the reader a complete picture of the results and to show that the upper bound is essentially tight.

Reviewer 2 also asks whether the kr/ϵ^2 upper bound for the non-efficient estimator's sample complexity exists in the literature, or may be easily obtained from [KOPS15], or from a sieve maximum likelihood estimator. We could not find any reference for the upper bound in the literature and would be happy to include one if pointed out. We also do not see how to obtain the result from [KOPS15], which does not leverage the structure of the distribution, hence will require k^2 samples, or from a sieve maximum likelihood estimator.

Reviewer 1 and 3 ask about learning in L_2 and spectral norms. Normalized L_1 norm is strictly stronger than normalized L_2 norm. Additionally, L_1 learning depends on the average row and column sums of M, yet as noted in [MD19] and mentioned in section 1.4, L_2 learning depends on the highest row and column sum, hence could suffer from a single heavy row. A similar observation about heavy rows and column sum holds for spectral norm as well. We will elaborate on this in the final version.

Reviewer 4 asks whether zeroing out rows and columns simultaneously could further reduce the algorithm's sample complexity. A small extension of Theorem 3 can show that difference matrix $X - M$ has a small submatrix (with few rows and columns) that when zeroed out will result in a remainder with a small spectral norm. Removing either the rows or the columns of this small submatrix would suffice, removing both would not yield significant additional advantage.

Reviewer 1 asks what are contaminated rows in Theorem 3. These are the rows we need to remove to achieve spectral concentration. Coincidentally, the same rows addressed in the previous paragraph. We will clarify that in the final version.

Reviewer 1 also asks whether statements in line 233 and 309 can be improved. Briefly, we believe the first can, but would require work. For the second, we do not know, but even if possible, it would improve the estimation error by at most a factor of two. We will also address typos and improve explanations as the reviewer suggests.

Finally, we would like to point out that although not addressed by the reviews, one consequence of our results, discussed in Section 1.4, significantly improves on state-of-the-art collaborative filtering results, e.g., [BCLS17].