We thank all the reviewers for their thorough and helpful remarks. We begin by addressing two main points, regarding theory for outlier detection (Reviewer 1) and additional experiments (Reviewers 2,3).

Theory for Outlier Detection We agree with Reviewer 1 that the outlier detection problems solved by our algorithm are exactly those where the outliers increase the operator norm of the empirical covariance from what it would have been with a clean version of the data set. (One might call them "spectral outliers".) These outlier detection problems are motivated in part by hard instances of the robust mean estimation problem – those on which algorithms preceding the 2016 works of Lai-Rao-Vempala and Diakonikolas-Kamath-Kane-Li-Moitra-Stewart would have incurred large errors. The example suggested by Reviewer 1, where the outlier distribution is a point-mass at the true mean of the inlier distribution, does not fit this spectral outliers setting. However, in this example, many statistics can be accurately estimated without removing the outliers (e.g. the empirical mean and covariance will be good estimators of the population mean and covariance, respectively).

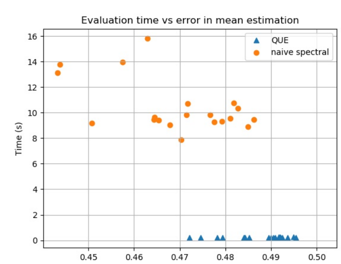
It is possible to formally state and prove such a result characterizing spectral outliers and our algorithm's success at such outlier detection tasks. Simple generative models (such as those we use to create synthetic data sets used in our experiments) lead to data sets with spectral outliers where our algorithm provably finds more outliers than e.g. naive spectral methods, outlier detection based on ℓ_2 norms, etc. (The proofs would use standard spectral analysis and concentration of measure – a small subset of the techniques used to prove our results on robust mean estimation.)

We elected not to formally define the task of outlier detection – we believe that any mathematical definition will fail to capture the breadth of real-world situations where outlier detection might be used. (By contrast, in the more theoretical portion of the paper on robust mean estimation, we have formal problem definitions and rigorous theorem statements.) That said, in any final version of our paper we will clarify with increased mathematical formality the subset of outlier detection tasks for which our outlier detection algorithm succeeds.

Robust Mean Estimation Experiments Reviewers 2 and 3 suggest additional experiments (1) comparing our algorithms against slower baselines for outlier detection, (2) measuring running time, and (3) on robust mean estimation in addition to outlier detection. We first note that we did compare against state-of-the-art methods based on *local outlier factors* – many of these algorithms do not run in linear time. See plots (g),(h),(i) and section 10 of supplementary material. The *naive spectral* algorithm which we compared against (captured by $\alpha \to \infty$ in our plots) is the natural outlier detection analogue of the slow-but-polynomial-time algorithms for robust mean estimation preceding our work – we observed experimentally that QUE scoring has improved AUC scores compared to this algorithm.

In preparing this author response we conducted some preliminary experiments on *robust mean estimation* using our synthetic data set. We compared two iterated filtering methods, one based on QUE scores and one based on naive spectral scores (the case of $\alpha \to \infty$) as used in prior work on polynomial-time algorithms for robust mean estimation (Figure). We iterated the filter, computing scores, removing a small fraction of the data points with the highest scores, and iterating until the empirical covariance had bounded spectral norm. We compared the accuracy (in Euclidean norm) and running times of both methods – both were implemented with optimized matrix multiplication libraries (BLAS) under PyTorch; we used our fast approximate QUE implementation (Section 11, supplement). To ensure a fair comparison, we did a hyperparameter sweep to optimize the runtime and accuracy of naive spectral filtering, and only included a single setting of hyperparameters for QUE scoring.

Our findings confirm the theory in our paper: although each iteration of QUE scoring is slightly slower than naive spectral scoring, the latter requires many more iterations to find a clean data set with bounded empirical covariance, resulting in much slower overall running time. For similar accuracy, the iterated filter algorithm with QUE scoring runs an order of magnitude faster than with naive spectral scoring.



Further Remarks We thank the reviewers for pointing out several typos – we will fix them. We agree with Reviewer 1 about the missing centering step, which we will fix. We will clarify the definitions of S_b, S_r and add a table of notation as suggested by Reviewer 3. We also agree about the error in Lemma 6.1, and with the suggested fix, which we will implement. Regarding Reviewer 3's question about equation (21): this was a typo (we thank you for pointing it out) – the correct step here is to bound the entire first term in the line above (b) by $\varepsilon\|M(w_t) - I\|\|\rho\|^2$. We have cleaned up this proof for future versions of the paper. If accepted, will increase the size of our plots using the additional page allowed in the camera-ready version.

Figure: We plot runtime vs mean estimation error (Euclidean norm) as fraction of points removed per iteration varies. We used synthetic data in 512 dimensions with 20% outliers spread across 20 orthogonal directions. Each data point represents a single run of iterated filtering. Removing fewer points per iteration leads to more iterations and thus slower running times; these are the runs with high running time but (slightly) better accuracy.