We thank the reviewers for their thoughtful reviews. Before we begin, we apologize for accidentally including the supplementary material in the main text. This was an honest mistake at submission time. We do not believe it changes anything in the review process, since the exact same supplementary material is included anyways, and we hope that the reviewers and ACs can overlook it.

General remarks We begin by addressing some shared questions from the reviewers.

Some of the reviewers were interested in seeing simulation studies for our algorithm. We agree with the reviewers that experimental evaluations would be interesting, and would be very useful follow-up. In our work, we focus on developing the theoretical foundations, which we believe are already valuable to the community. However, we do believe our algorithm is eminently practical, as it does not require any heavy-duty theoretical tools. In contrast, the algorithm in [CDGW19] is quite complex, and uses some complex optimization techniques which are likely impractical.

Some of the reviewers also had questions about the relationship of our work with previous ones, specifically [CDGW19] and [DHL19]. We will clarify the relationship more thoroughly in the next version of the paper, and we will also briefly explain here. At a very high level, we follow the same algorithm structure as [CDGW19], e.g. in Alg 1, and we use their subroutine to get a crude estimate of the covariance, e.g. in Alg 2. However, most of the technical work in both papers lies within the fine-tuning step (Alg 3 in our paper), and here we use very different tools. [CDGW19] relies on very heavy-duty optimization primitives, specifically, nearly-linear time packing SDP solvers. In contrast, we design much more streamlined and specialized tools to solve our optimization problems, which exploit the structure of the problem much more effectively. Morally, this is why we are able to improve the runtime of [CDGW19].

We do directly build off of the framework in [DHL19]. However, there are a number of non-trivial technical challenges to adapt their framework—which was developed for mean estimation—to the covariance estimation setup, which we must overcome. We also believe that it is an interesting contribution in its own right to demonstrate the generality of the QUE scoring framework. In particular, we found it somewhat surprising that it is able to produce arguably one of the cleanest algorithms for such a fundamental problem in robust statistics, which is simultaneously optimal in terms of runtime, sample complexity, and error (up to logarithmic factors).

Reviewer specific comments We thank all the reviewers for detailed comments, and we will fix the minor issues the reviewers pointed out in the next version. We now address individual reviewer's questions and comments that were not addressed above.

Reviewer 1: The log κ comes from our invocation of of Lemma 3.1 in [CDGW19], which is used to get a rough estimate the covariance. Intuitively, this log κ comes from the fact that first order methods such as matrix multiplicative weights typically must pay a logarithmic factor in the "width" of the problem, and the only width bound we can typically obtain a priori depends on κ. It is an interesting theoretical question to avoid paying this factor with linear-time algorithms (note that slower polynomial time algorithms do not have to pay this factor). We believe this is a mild cost, since we only depend logarithmically on κ, and in almost all applications, the condition number is polynomially bounded.

Reviewer 2: We thank the reviewer for their kind words. Please see discussion above related to their comments regarding simulation and comparison to prior work.

Reviewer 3:

- As briefly mentioned above, the speed-up comes from designing better optimization routines that heavily utilize the structure of the problem. At a very high level, while the width of the problem does scale with ε, we demonstrate that the uncorrupted data satisfies the regularity condition (Def. 3.2) actually allows us to pretend that the width is constant, even when we are running matrix multiplicative weights to high precision. We will explain this more thoroughly in the next version of the paper.
- Def 3.2 says that not only do the empirical mean and covariance of the data concentrate, but any large subset of the data still has good concentration. Intuitively, we should expect this to be true because Gaussians exhibit very strong concentration, and this still holds when we condition on events with probability $1 - \varepsilon$.

Reviewer 4:

- As discussed in lines 127-130, our results hold (at a cost of doubling ε) for Gaussians with arbitrary mean. As mentioned in lines 58-59, by combining this with previous work, this allows us to learn arbitrary Gaussians to good total variation distance.
- Please see previous discussion in regards to the relationship to [CDGW19].
- We agree that the approximate score oracles are important and we discuss them at a high level in Sec. 3.3.2. However, the details are quite technical, and due to space considerations we chose to move the formal discussion to the supplementary material, in favor of presenting (what we consider to be) the main takeaways in the main text.