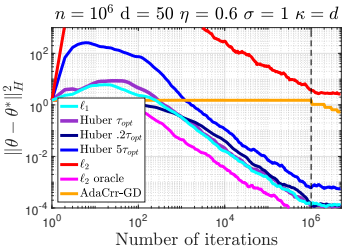
We thank all the reviewers for their feedback and comments. They will be incorporated into the revised version.

We would like to highlight several points regarding our contribution. In the paper we consider the robust linear regression problem which has drawn significant interest in the last decades. Even in the simpler setting where only the responses are obliviously corrupted there have been numerous works proposing novel algorithms that recover the gold model θ^*. Though this setting may seem simple, it has considerable applications. Nonetheless, currently none of the proposed algorithms are suitable for modern large-scale problems. This observation is the incentive behind our work: we propose the first highly scalable and efficient algorithm for the robust linear regression problem with oblivious response corruptions. We show that running SGD on the ℓ_1 loss outperforms all current algorithms, theoretically and empirically. Not only is it currently the unique suitable large-scale algorithm, but it also presents several major advantages even in the offline setting: a) there are no hyperparameters to tune, the knowledge of σ or η are not necessary b) the $O(1/n)$ dominant convergence terms are independent of the conditioning κ of the data c) the algorithm adapts to the difficulty of the adversary since the rate depends on the effective corruption proportion $\tilde{\eta}$. We also want to underline that the algorithm is extremely simple to implement: it represents only two lines of code.

Concerning the Gaussian assumption on the data, we agree that it is restrictive and that it does not always represent real world situations. However we highlight that this assumption (or similar specific structural assumptions on the design matrix and the noise) is made in the previous papers that deal with the framework that we consider. The key for our analysis is that the continuous density of (x, ε) smooths the objective f, this enables a proof of type Polyak-Judistky to obtain the final rate. The more restrictive Gaussian assumption is made in order to obtain a closed form for f which enables simpler (but not simple !) computations to obtain the final convergence rate. We believe that the proof could be extended to any sub-Gaussian continuous data, however this would require considerably heavier and cumbersome computations (see the figure for an example where the covariates are sampled from a zero-mean uniform distribution with covariance H and the noise from a uniform distribution with variance σ^2).

To Reviewer 1. As pointed out, the dependency on η (or rather $\tilde{\eta}$) we obtain might not be optimal and is still an interesting open question. It is however the best dependency which has been obtained since the effective outlier proportion $\tilde{\eta}$ is strictly smaller than η. We agree that real data experiments would have been a plus, they will be added to the revised version.



Online robust regression on non-Gaussian data. Same adversarial setting as in Figure 1 in the paper.

To Reviewer 2. The centered data assumption is a classical assumption which is often made in machine learning, note that it is for instance made in previous papers that deal with the framework we consider. This assumption is reasonable since it is of common practice to pre-process and center the features. We highlight the fact that even for the Gaussian linear regression problem the minimax rate of estimation is $O(1/n)$, in the sense that no estimator that uses the same information can improve upon this rate. Therefore, in the setting we consider, the rate we obtain cannot be improved.

To Reviewer 3. The linear model is one of the simplest model we could have considered and linear regression is certainly amongst the oldest and most fundamental statistical methods. However: (a) it is still intensively used in practice and studied in theory, (b) more complex models can be seen as an extension of it and (c) there are still many interesting open questions. Indeed even for this simple model there were still no efficient algorithms which could deal with large datasets, hence the motivation of our work. The extension of the algorithm to broader and more general models is undoubtedly interesting, however it is outside the scope of our present work.

To Reviewer 4. As seen in Theorem 4, the terms depending on the variance σ go to 0 as $\sigma \to 0$. Hence the extreme case where there is no 'nice' noise is not pathological and the algorithm still performs well. On the other hand, the performance of the algorithm is indeed inevitably dependent of $\tilde{\eta}$ (and not on η as for other robust algorithms): we cannot expect any algorithm to behave well as $\tilde{\eta}$ goes to 1 since this corresponds to having all the data corrupted. The dependency $(1 - \tilde{\eta})^{-2}$ we obtain is to our knowledge the best which has been obtained.

We thank the reviewer for the interesting missing reference which will be added to the revised version. The literature on robust regression is vast and we have tried to be as exhaustive as possible. The mentioned paper proposes an interesting EM-algorithm, however they do not give any convergence guarantee. Furthermore every M-step of their algorithm is computationally heavy since it requires solving a weighted LS problem on the whole dataset, making it inappropriate for large-scale problems.

As pointed out, the Tukey loss is often used to deal with corruptions, however its non-convexity makes it much harder to analyse and optimise. Furthermore, as for the Huber loss, the Tukey loss requires tuning an extra parameter and we believe it will not lead to better experimental results for the linear model with corrupted response. However the Tukey loss (as well as other redescending M-estimators) is well-known for being more robust to corruption in the features x. We let further investigations as future work.