We thank all the reviewers for the helpful comments. Here, we address the main concerns raised by the reviewers.

To Reviewer # 3 [1.1. Motivations for regression with sparse interaction terms.] Regression with interaction terms has been studied for a long time (see [1] and its citations). The main obstacle comes from the dramatically increased dimensionality due to considering the quadratic or even higher order interaction terms. An immediate remedy is adding constraints to the model, which helps to reduce the model complexity. In our work, we adopted the sparse constraint, for both computational efficiency and interpretability.

[1.2. Real application with sparse Θ] The sparse interaction assumption also holds in many real applications. One example comes from genome-wide association studies. For a given phenotype, the associated genetic variants are usually a sparse subset of all possible variants. While the traditional method (e.g., Lasso) can find important individual genes, our method is able to find the sparse interaction between two (and potentially multiple) genes, which is especially desirable based on the biological knowledge that genes work together in structured groups [2].

[1.3. Novelty of our work] The key new realization of our paper is that, for our specific problem, it is possible to find the top 2k elements of the gradient without even calculating the entire gradient. This allows for our method, which is iterative hard thresholding (IHT) with approximate support recovery via count sketches, to run in sub-quadratic time.

The existing analyses of IHT does not carry over to our setting - because we have inexactness in finding the top-k AND inexactness in finding the gradient. Thus, our analysis has two main contributions: (1) showing that our count sketch based support recovery gives good approximation guarantees under the assumptions of restricted smoothness and sparsity of the optimal solution, and (2) showing that our IHT variant still converges linearly under this inexact support recovery. Thus our paper bridges the (so far separate) analyses of count sketch and stochastic iterative hard thresholding, for an important use case of finding higher order interactions.

To Reviewer # 4 [2.1. Motivation.] The motivations and application are discussed in [1.1.] and [1.2.].

[2.2. Optimality of the current result.] $O(k \log p)$ samples, along with our landscape assumptions of restricted smoothness etc, are known information theoretic lower bounds for recovering a k-sparse $\mathbf{\Theta}$. By Theorem 5, our method matches this lower bound up to a constant factor. Similarly, for the time and space complexity, the optimal complexity is $\Omega(kp)$, since a minimum of $\Omega(k)$ samples are required for recovery, and $\Omega(p)$ for reading all the entries. Corollary 6 shows that the time and space complexity of our method is $\widetilde{O}(k(k+p))$, which is near optimal. These results are briefly mentioned in the paper, we would like to highlight them in the revised version.

[2.3. More explanation for the statistical and computational trade-off.] In our method, the parameter b (which controls the output size of the count sketch) determines the statistical and computational trade-off. In Figure 1-(a), the black dashed line stands for solving the quadratic regression with exact gradient calculation, which is a statistical benchmark (not achievable in sub-quadratic time). By choosing larger b, we are getting closer to the that while paying extra computation (recall that the computational complexity is given by $\widetilde{O}(m(p+b))$, where m is the batch size and p is the dimension of input feature $\mathbf{x}$). As shown in Theorem 1 and Lemma 2, setting b to the same order as sparsity k is sufficient for the consistent parameter estimation. Setting b to p^2 will yield exact gradient calculation while incurring quadratic complexity.

To Reviewer # 5 [3.1. When the sparse assumption doesn't hold] Theoretically, the sparsity assumption is commonly adopted in high dimensional statistics. In the case when the ground truth is dense or the sparsity parameter k is set to be smaller than the true sparsity K, some preliminary experiments indicate that both our method, and more classical ones like Lasso/standard IHT, can converge to a poor sparse solutions - unless there are some other extraneous assumptions. Thus lack of underlying sparsity in the true $\mathbf{\Theta}^*$ is a problem for all sparse recovery methods.

Note that Corollary 6 shows that the overall time complexity is $\widetilde{O}(k(k+p))$, where p is the dimension of input feature $\mathbf{x}$. A lower bound on the complexity is $\Omega(kp)$. Thus any method can have sub-quadratic complexity only when the sparsity k of the truth is smaller than $O(p)$; this is both necessary and sufficient for our model as well.

[3.2. Motivations] The motivations and one real application where sparsity holds are discussed in [1.1.] and [1.2.].

References

[1] James Jaccard and Robert Turrisi. *Interaction effects in multiple regression*, volume 72. Sage, 2003.

[2] Yun Li, George T O'Connor, Josée Dupuis, and Eric Kolaczyk. Modeling gene-covariate interactions in sparse regression with group structure for genome-wide association studies. *Statistical applications in genetics and molecular biology*, 14(3):265–277, 2015.