To Reviewer 1 and 3: We are very very fortunate to receive such detailed suggestions that we can bound the nuclear form of the variance and analyze the condition under which the variance of LGD would be better than that of SGD from the reviewer who fully understands our work!! The direction the reviewer pointed to is absolutely the right way to do the analysis, including bounding random variable |S| which itself in expectation is a summation of all collision probabilities. The analysis for bounding variance is still little involved and will require tools from [1] (like holders inequality), which we cited at line 235 in our paper. We will add a discussion in the subsequent version of the paper.

A sneak peek of the summary: Our LGD can be viewed as a Kernel Density Estimation(KDE) defined in Equation (1) from citation [1]. From Lemma (3) of [1], we can similarly get rid of our |S| term by applying Bayes rule to condition on the random variable $|S|$. Let $p_1 \geq p_2 \geq ... \geq p_N$,

$$\frac{1}{N^2}(\sum_{i=1}^{N_1} \frac{\|\nabla f(x_i, \theta_t)\|_2^2 \cdot |S|}{p_i} - \frac{1}{N^2}\|(\sum_{i=1}^{N} \nabla f(x_i, \theta_t))\|_2^2 \leq \frac{1}{N^2}(\sum_{i=1}^{N} \frac{\|\nabla f(x_i, \theta_t)\|_2^2}{p_i}(i + \sum_{j=i+1}^{N} \frac{p_j}{p_i}) - \frac{1}{N^2}\|(\sum_{i=1}^{N} \nabla f(x_i, \theta_t))\|_2^2.$$

We denoted the expectation of SGD or LGD as μ. [1] first showED that the trace of the covariance of SGD is tight up to constants in the worst case, $\mu^2 * \mu^{-1} - \frac{1}{N^2}\|(\sum_{i=1}^{N} \nabla f(x_i, \theta_t))\|_2^2$. The definition of (β, M)-scale free estimator is in Definition(3) and that of (τ, γ)-localized query is in Definition(5) of the paper [1]. We will then see that the upper bound on the variance improves when most of the contribution to μ comes from relatively large gradient norms, which gives an intuition about when the variance is better. Let LGD be a (β, M)-scale free estimator with $\beta \in [1/2, 1]$. For every (τ, γ)-localized query θ_t, the upper bound of the trace of the covariance of LGD estimator would be,

$$Tr(\Sigma(Est)) \leq \mu^2 * M^3\{2\gamma^\beta + \gamma^{2-\beta} + \tau^{2\beta-1}\gamma^\beta\}\mu^{-\beta} - \frac{1}{N^2}\|(\sum_{i=1}^{N} \nabla f(x_i, \theta_t))\|_2^2. \quad (1)$$

It also pointed out if there is no assumption made ($\gamma = \tau = 1$), the bound becomes, $\mu^2 * 4M^3\mu^{-\beta} - \frac{1}{N^2}\|(\sum_{i=1}^{N} \nabla f(x_i, \theta_t))\|_2^2$. where the optimal choice of β is $\frac{1}{2}$.

We agree with the reviewer that the sampling version of DCI can also be derived. We do not need to aggregate the samples from all indices to find the nearest neighbor but may just sample several ones similar to LGD. We can easily add this as a comparison in our paper if needed. We want to point the reviewer to page 5 of our paper that we did provide empirical analysis and discussions on the change of per-iteration cost, running time and N, K, L, d. We are also happy to add more thorough experiments for analyzing the effectiveness of all these parameters in the main paper. We chose Simhash because cosine similarity is useful for the task of inner product of two vectors but we will add a discussion on choosing teh hash family in the paper.

To Reviewer 2: We thank the reviewer for explicitly raising several concerns. However, they are not related to the paper. Since we think the reviewer has many misunderstanding of the main argument of the paper, we will explain it here again, and we hope to help the reviewer clarify our goal. The goal of the paper is to provide a better gradient estimation procedure rather than improving the training of a certain model. SGD is indeed the right baselines as there is no faster way, in terms of computational cost, to estimate the gradient than random sampling, as mentioned in the introduction. We have derived the estimation for linear regression, logistic regression, and even neural networks. The experiment on Bert is not about improving the training accuracy but to show the effectiveness of our superior and fast estimation of the gradient for faster convergence. Besides, we carefully read both papers referred to in the review, and we want to clarify that we are not doing near neighbor search. There is **no similarity search with LSH** in our proposal; it is sampling and unbiased estimation.

We sincerely hope the restatement will help the reviewers become clear about the challenges, motivation, and the goal of our paper and potentially change the negative view formed due to misunderstanding.

To Reviewer 3: We are delighted to see the reviewer's interest in our idea! We want to do a few clarifications, and we hope the reviewer can have a better understanding of our algorithm and the contribution of the paper. We propose the first algorithm that achieves sharper estimates of the gradient in near-constant time using hash tables. As a result, we speed up any first-order gradient-based algorithm, including adagrad, SGD, Adam, etc. Our algorithm uses LSH to do adaptive sampling, and it is not only restricted to regression tasks. We agree that only for linear models, we can show connections with optimal sampling. However, for the sampling to be better than random (current practice), we only want positive correlation with L_2 norm of gradient. We can achieve that even by linearizing any non-linear model. Thus, even for neural networks, we have an informed sampling (it is not optimal, but still better than random sample based gradient estimation.). We did show the experiments on NN in Section 3.2.

We thank the reviewer for pointing out the problems in our supplementary and give us a chance for a likely score increase. We apologize for the mislabelling. It should be (11)-(7), (17)->(9), (16)->(8). We will address the typo and other problems in the main paper and appendix.

[1] Charikar, Moses and Siminelakis, Paris. *Hashing-based-estimators for kernel density in high dimensions* 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), 2017.